

Appendix for Reducing Disparity Between LLMs and Humans: Optimal LLM Sample Calibration

A Theoretical Results and Proofs	A-2
A.1 Theorem A-1 (Convex Minimizer)	A-2
A.2 Proposition A-2 (HuMCal Weakly Dominates Any Feasible Fixed Weighting) and Corollary A-3 (Baseline-Safe)	A-4
A.3 Proposition A-4 (Scalar Finite-Sample Oracle Inequality for Simplex ERM)	A-4
A.4 Theorem A-6 (Near-Optimal Transfer Under Moderated Domain Shift)	A-7
A.5 Proposition A-7 (Monotonic Improvement Under Uniform Upgrades)	A-9
A.6 Finite-Step Softmax Implementation	A-10
B Intuition Through Hypothetical Numerical and Geometric Illustrations	A-11
B.1 Hypothetical Numerical Illustration of the Mechanism	A-11
B.2 Geometric Explanation of HuMCal	A-11
C Out-of-Sample Sensitivity Analyses and Subgroup-Design Diagnostics	A-13
C.1 A Simple Decomposition of Out-of-Sample Transfer Performance	A-13
C.2 Explicit Regularization: Limited Gains beyond the Simplex Constraint	A-14
C.3 Finite-Sample Robustness: Diminishing Returns and No-Signal Candidates	A-14
C.4 Regression Diagnostic: How Subgroup Characteristics Relate to Marginal Gains	A-16
C.5 Weight Diagnostics	A-18
C.6 Practical Subgroup Design: Open-Ended vs. Menu-Based Tasks	A-20
C.7 Checkpoint-Upgrade Evaluation (Model-Improvement Compatibility)	A-21
D Market-Structure Data and Supplementary Analyses	A-23
D.1 LLM Data Collection	A-23
D.2 Processing Human Behavioral Data	A-24
D.3 Objective Function for Regional Market Structure (Ordinal Co-Purchase Data)	A-25
D.4 Calibration Weights for Market Structure Learning	A-26
D.5 Agreement Rate (AR) for Market Structure Learning	A-27
D.6 Hierarchical Calibration with Subgroup-Level Benchmarks	A-29
D.7 Does HuMCal Still Help When No Persona Beats the Aggregate Prompt?	A-30
E Public Opinion Data and Supplementary Analyses	A-32
E.1 Data Processing on Public Opinion Surveys	A-32
E.2 Objective Function for Public-Opinion Learning (Categorical Data)	A-36
E.3 Calibrated Weights over Checkpoint–Persona Subgroups	A-37
E.4 Supplementary Analyses Using Wasserstein Distance	A-37

Appendix A: Theoretical Results and Proofs

A.1. Theorem A-1 (Convex Minimizer)

THEOREM A-1 (Convex Minimizer). Assume (i) each mapping $\mathbf{w} \mapsto \mathcal{Z}_q^L(\mathbf{w})$ is affine in $\mathbf{w}$, and (ii) for each question q , the map $z \mapsto \mathcal{D}(z, \mathcal{Z}_q^H)$ is convex and continuous. Then the optimization (HuMCal-OPT) is a convex program over the simplex Δ_G . Consequently, a global minimizer exists, every local minimizer is global, and any feasible point satisfying the Karush–Kuhn–Tucker (KKT) subgradient conditions is a global minimizer. In particular, HuMCal attains

$$\min_{\mathbf{w} \in \Delta_G} \sum_{q=1}^Q \mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H).$$

If, in addition, the composed map

$$\mathbf{w} \mapsto \mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H)$$

is strictly convex on Δ_G for at least one q , then the full objective is strictly convex on Δ_G and the minimizer is unique.

Uniqueness of the weights. Without strict convexity the minimizer need not be unique: if two subgroups produce identical outputs on every calibration item, any positive combined weight assigned to them can be redistributed without changing the fitted aggregate or the calibration loss. Step 7 reuses the weights themselves, so source-equivalent weightings can diverge on a target domain.

Proof of Theorem A-1.

Step 1: Convex objective. For each question q , the map $\mathbf{w} \mapsto \mathcal{Z}_q^L(\mathbf{w})$ is affine, and $z \mapsto \mathcal{D}(z, \mathcal{Z}_q^H)$ is convex and continuous. Hence

$$f(\mathbf{w}) = \sum_{q=1}^Q \mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H)$$

is convex in $\mathbf{w}$. Because affine maps are continuous, and continuity is preserved under composition and finite sums, f is also continuous on Δ_G .

Step 2: Convex feasible set (the simplex). The feasible set

$$\Delta_G := \{\mathbf{w} \in \mathbb{R}^G : w_g \geq 0 \ \forall g, \sum_{g=1}^G w_g = 1\}$$

is an intersection of the non-negative orthant and an affine hyperplane, so it is convex, closed, and bounded. Because f is continuous and Δ_G is compact, a minimizer exists (Weierstrass theorem).

Step 3: No spurious local minima. The program $\min_{\mathbf{w} \in \Delta_G} f(\mathbf{w})$ is a convex optimization problem. Therefore, every local minimizer is necessarily global.

Step 4: subgradient-KKT conditions are sufficient. Let $\mathbf{w}^* \in \Delta_G$ satisfy the subgradient-KKT conditions. Then there exist a subgradient $\mathbf{s} \in \partial f(\mathbf{w}^*)$, multipliers $\boldsymbol{\mu} \in \mathbb{R}_+^G$ for the constraints $-w_g \leq 0$, and a multiplier $\nu \in \mathbb{R}$ for $\mathbf{1}^\top \mathbf{w} = 1$, such that

$$\mathbf{s} - \boldsymbol{\mu} + \nu \mathbf{1} = \mathbf{0}, \quad \mu_g w_g^* = 0 \quad \text{for every } g.$$

For any $\mathbf{v} \in \Delta_G$, the subgradient inequality gives

$$\begin{aligned} f(\mathbf{v}) &\geq f(\mathbf{w}^*) + \mathbf{s}^\top (\mathbf{v} - \mathbf{w}^*) \\ &= f(\mathbf{w}^*) + \boldsymbol{\mu}^\top \mathbf{v} \geq f(\mathbf{w}^*). \end{aligned}$$

The equality uses $\mathbf{1}^\top \mathbf{v} = \mathbf{1}^\top \mathbf{w}^* = 1$ and complementary slackness; the final inequality uses $\boldsymbol{\mu} \geq \mathbf{0}$ and $\mathbf{v} \geq \mathbf{0}$. Thus $\mathbf{w}^*$ is a global minimizer.

Step 5: Statement of the result. Thus $\mathbf{w}^*$ attains the smallest possible total discrepancy

$$\sum_{q=1}^Q \mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}^*), \mathcal{Z}_q^H),$$

over the simplex Δ_G . Equivalently, no feasible weight vector yields a smaller summed discrepancy across the calibration questions under the chosen distance $\mathcal{D}$.²³

If, for at least one q , the composed map

$$\mathbf{w} \mapsto \mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H)$$

is strictly convex on Δ_G , then f is strictly convex and the minimizer is unique. Without strict convexity of the composed objective, uniqueness need not hold. $\square$

²³ This is a statement about the summed objective across calibration questions. The aggregate need not be individually closest to $\mathcal{Z}_q^H$ for every question q .

A.2. Proposition A-2 (HuMCal Weakly Dominates Any Feasible Fixed Weighting) and Corollary A-3 (Baseline-Safe)

Because an exact HuMCal solution globally minimizes the calibration objective over a feasible set containing each comparator, it cannot have greater calibration loss than any feasible comparator weighting. The following proposition formalizes this.

PROPOSITION A-2 (HuMCal Weakly Dominates Any Feasible Fixed Weighting). *Under the conditions of Theorem A-1, let $f(\mathbf{w}) = \sum_{q=1}^Q \mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H)$ be the HuMCal objective. Let $\Delta_G = \{\mathbf{w} : w_g \geq 0, \sum_g w_g = 1\}$ be the probability simplex. If $\mathbf{w}^*$ is any global minimizer of f over Δ_G , then*

$$f(\mathbf{w}^*) \leq f(\mathbf{v}) \quad \text{for all } \mathbf{v} \in \Delta_G.$$

The inequality is strict whenever $\mathbf{v}$ is not itself a global minimizer.

Proof of Proposition A-2. By the definition of a global minimizer, $f(\mathbf{w}^*) = \min_{\mathbf{w} \in \Delta_G} f(\mathbf{w})$. Hence $f(\mathbf{w}^*) \leq f(\mathbf{v})$ for every $\mathbf{v} \in \Delta_G$. If $\mathbf{v}$ is not a global minimizer, then $f(\mathbf{v}) > f(\mathbf{w}^*)$, so the inequality is strict. $\square$

Common fixed-weighting heuristics correspond, once specified or selected, to feasible simplex vectors shared across the calibration items. The best-single vector is selected on the calibration split rather than predetermined. HuMCal, by contrast, globally minimizes the calibration objective over that simplex, so its exact solution cannot underperform any such feasible comparator on the calibration split. The following corollary formalizes this point.

COROLLARY A-3 (Baseline-Safe). *Under the conditions of Proposition A-2, HuMCal’s optimizer $\mathbf{w}^*$ weakly dominates every comparator vector $\mathbf{v} \in \Delta_G$, whether prespecified or selected on the calibration split. In particular, it dominates:*

- (a) *population weighting (demographic post-stratification): when the subgroup set is demographic, $\pi = (\pi_1, \dots, \pi_G)$ assigns each subgroup its census share (with zero weight on any included non-demographic subgroup, such as the aggregate prompt, so that the shares sum to one).*
- (b) *direct elicitation (single aggregate prompt): $\mathbf{e}_d$ places all weight on the included aggregate subgroup d .*
- (c) *uniform weighting across model variants: $\mathbf{u} = (1/G, \dots, 1/G)$ assigns equal mass to every LLM check-point or variant.*
- (d) *best single subgroup: any simplex vertex $\mathbf{e}_g$, including the best single subgroup selected on the calibration split.*

For each feasible baseline vector $\mathbf{v}$ in this list, we have the weak inequality $f(\mathbf{w}^) \leq f(\mathbf{v})$, with strict inequality whenever $\mathbf{v}$ is not itself a global minimizer.*

A.3. Proposition A-4 (Scalar Finite-Sample Oracle Inequality for Simplex ERM)

This subsection gives stylized finite-sample guidance under an idealized model of i.i.d. complete observations. Proposition A-4 bounds population-risk generalization for exact simplex ERM; we then state conditions for extending its $\sqrt{\log G/Q}$ dependence to the objectives used in the two applications.

Setup and assumptions. Let $q = 1, \dots, Q$ index calibration items (e.g., brands or survey questions). For each item q and subgroup $g \in \{1, \dots, G\}$, let $Z_{q,g}^L \in \mathbb{R}$ denote a scalar subgroup output and let $Z_q^H \in \mathbb{R}$ denote the corresponding scalar human benchmark. For any $\mathbf{w} \in \Delta_G$, define the aggregated prediction $Z_q^L(\mathbf{w}) = \sum_{g=1}^G w_g Z_{q,g}^L$. Let $\ell(\cdot, \cdot)$ be the per-item loss (induced by $\mathcal{D}$ in the main text) and define $\ell_q(\mathbf{w}) = \ell(Z_q^L(\mathbf{w}), Z_q^H)$. Define the population risk $R(\mathbf{w}) = \mathbb{E}[\ell_q(\mathbf{w})]$ and empirical risk $\hat{R}(\mathbf{w}) = \frac{1}{Q} \sum_{q=1}^Q \ell_q(\mathbf{w})$. Let $\hat{\mathbf{w}} \in \arg \min_{\mathbf{w} \in \Delta_G} \hat{R}(\mathbf{w})$ denote a HuMCal empirical risk minimization (ERM) solution. We make the following assumptions:

- (A0) (i.i.d. calibration observations)** The complete item-level observation vectors $X_q := (Z_q^H, Z_{q,1}^L, \dots, Z_{q,G}^L)$, $q = 1, \dots, Q$, are i.i.d. draws from a distribution $\mathcal{P}$. The candidate set, subgroup-output construction, and loss are either fixed independently of the calibration sample or treated as fixed after conditioning on independent upstream data. The components of X_q may be arbitrarily dependent within an item.
- (A1) (Bounded loss)** $0 \leq \ell_q(\mathbf{w}) \leq B_\ell$ for all q and all $\mathbf{w} \in \Delta_G$.
- (A2) (Lipschitz in prediction)** For every z^H , the map $z \mapsto \ell(z, z^H)$ is L -Lipschitz: $|\ell(z, z^H) - \ell(z', z^H)| \leq L|z - z'|$.
- (A3) (Bounded subgroup outputs)** $|Z_{q,g}^L| \leq B$ for all q, g .

Assumption **(A0)** treats complete calibration observations as independent draws from a common distribution. Yet, the focal brands in the market-structure application share a transaction network, and the Pew items form a curated question inventory. We therefore use Proposition A-4 to characterize how the estimation term depends on G and Q ; its probability statement applies to the idealized i.i.d. setting. The result treats Z_q^H as observed and does not propagate uncertainty from estimating the human benchmark.

PROPOSITION A-4 (Scalar Finite-Sample Oracle Inequality for Simplex ERM). *Under assumptions **(A0)**–**(A3)**, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of the complete calibration observations $X_1, \dots, X_Q$,*

$$\sup_{\mathbf{w} \in \Delta_G} |R(\mathbf{w}) - \hat{R}(\mathbf{w})| \leq 2LB \sqrt{\frac{2 \log G}{Q}} + B_\ell \sqrt{\frac{\log(2/\delta)}{2Q}}. \quad (\text{A.1})$$

Consequently, the ERM solution $\hat{\mathbf{w}}$ satisfies the oracle inequality

$$R(\hat{\mathbf{w}}) - \min_{\mathbf{w} \in \Delta_G} R(\mathbf{w}) \leq 4LB \sqrt{\frac{2 \log G}{Q}} + 2B_\ell \sqrt{\frac{\log(2/\delta)}{2Q}}. \quad (\text{A.2})$$

Holding L , B , and B_ℓ fixed, the displayed estimation term depends on the number of subgroups only through $\sqrt{\log G}$.

The bound gives an estimation term of order $\sqrt{\log G/Q}$. Enlarging the candidate set therefore increases the worst-case estimation cost slowly in G . The shapes of the observed learning curves are empirical.²⁴

²⁴ The constants follow the standard symmetrization, contraction, and Massart arguments and are not optimized, so the inequality should be read as an upper bound on the estimation term. The order $\sqrt{\log G/Q}$ is the relevant rate for ERM over a simplex with G vertices. Under additional low-noise conditions, localized analyses can deliver faster rates.

COROLLARY A-5 (Comparison to the Best Single Subgroup). *Under the conditions of Proposition A-4, since each vertex $\mathbf{e}_g$ is feasible in Δ_G , we also have with probability at least $1 - \delta$ over the draw of the Q calibration items,*

$$R(\hat{\mathbf{w}}) - \min_{g \in \{1, \dots, G\}} R(\mathbf{e}_g) \leq 4LB \sqrt{\frac{2 \log G}{Q}} + 2B_\ell \sqrt{\frac{\log(2/\delta)}{2Q}}. \quad (\text{A.3})$$

Thus, whenever the best convex mixture improves upon the best single subgroup by more than the right-hand side, *HuMCal* also improves upon the best-single-subgroup benchmark in population risk.

Proof of Proposition A-4.

We use a standard symmetrization/contraction argument based on expected Rademacher complexity.

Step 1: Function classes and Rademacher complexity. Let $q_1, \dots, q_Q$ denote the sampled complete observations $X_1, \dots, X_Q$ from assumption **(A0)**, and let $\sigma_1, \dots, \sigma_Q$ be i.i.d. Rademacher random variables taking values in $\{+1, -1\}$. For a class of functions $\mathcal{F}$ mapping complete observations to real numbers, define the expected Rademacher complexity at sample size Q by

$$\mathfrak{R}_Q(\mathcal{F}) := \mathbb{E}_{q_1, \dots, q_Q, \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{Q} \sum_{i=1}^Q \sigma_i f(q_i) \right].$$

Consider the prediction class

$$\mathcal{H}_G := \left\{ q \mapsto Z_q^L(\mathbf{w}) = \sum_{g=1}^G w_g Z_{q,g}^L : \mathbf{w} \in \Delta_G \right\}$$

and the induced loss class

$$\mathcal{L}_G := \{ q \mapsto \ell(Z_q^L(\mathbf{w}), Z_q^H) : \mathbf{w} \in \Delta_G \}.$$

Step 2: Contraction for Lipschitz loss. For a realized sample, define $\phi_i(z) := \ell(z, Z_{q_i}^H) - \ell(0, Z_{q_i}^H)$ for $i = 1, \dots, Q$. Each ϕ_i is L -Lipschitz and satisfies $\phi_i(0) = 0$. The removed term $Q^{-1} \sum_i \sigma_i \ell(0, Z_{q_i}^H)$ does not depend on $\mathbf{w}$, can be taken outside the supremum, and has conditional expectation zero over the Rademacher signs. By the coordinatewise Ledoux–Talagrand contraction inequality and assumption **(A2)**,

$$\mathfrak{R}_Q(\mathcal{L}_G) \leq L \mathfrak{R}_Q(\mathcal{H}_G). \quad (\text{A.4})$$

Step 3: Rademacher complexity of simplex mixtures (convex hull of G experts). For any realized sample $(q_1, \dots, q_Q)$ and any fixed Rademacher draw σ , we have

$$\sup_{\mathbf{w} \in \Delta_G} \sum_{i=1}^Q \sigma_i Z_{q_i}^L(\mathbf{w}) = \sup_{\mathbf{w} \in \Delta_G} \sum_{g=1}^G w_g \left(\sum_{i=1}^Q \sigma_i Z_{q_i,g}^L \right) = \max_{g \in \{1, \dots, G\}} \sum_{i=1}^Q \sigma_i Z_{q_i,g}^L,$$

since a linear objective over the simplex is maximized at an extreme point. Therefore,

$$\mathfrak{R}_Q(\mathcal{H}_G) = \mathbb{E}_{q_1, \dots, q_Q, \sigma} \left[\max_{g \in \{1, \dots, G\}} \frac{1}{Q} \sum_{i=1}^Q \sigma_i Z_{q_i,g}^L \right]. \quad (\text{A.5})$$

By Massart’s finite-class lemma, conditional on the realized sample,

$$\mathbb{E}_\sigma \left[\max_g \sum_{i=1}^Q \sigma_i Z_{q_i,g}^L \middle| q_1, \dots, q_Q \right] \leq \sqrt{2 \log G} \max_g \left(\sum_{i=1}^Q (Z_{q_i,g}^L)^2 \right)^{1/2}.$$

Using assumption **(A3)**, $\sum_{i=1}^Q (Z_{q_i,g}^L)^2 \leq QB^2$ for each g , so after dividing by Q and taking expectation over the sample,

$$\mathfrak{R}_Q(\mathcal{H}_G) \leq B \sqrt{\frac{2 \log G}{Q}}.$$

Combining with (A.4) yields

$$\mathfrak{R}_Q(\mathcal{L}_G) \leq LB \sqrt{\frac{2 \log G}{Q}}. \quad (\text{A.6})$$

Step 4: Uniform deviation bound. Apply the one-sided expected-Rademacher deviation bound to $\mathcal{L}_G$ and $-\mathcal{L}_G$, then use symmetry and a union bound. By assumption **(A1)**, with probability at least $1 - \delta$,

$$\sup_{\mathbf{w} \in \Delta_G} |R(\mathbf{w}) - \hat{R}(\mathbf{w})| \leq 2\mathfrak{R}_Q(\mathcal{L}_G) + B_\ell \sqrt{\frac{\log(2/\delta)}{2Q}}.$$

Substituting (A.6) gives (A.1).

Step 5: Oracle inequality for ERM. Finally, the standard ERM inequality implies

$$R(\hat{\mathbf{w}}) - \min_{\mathbf{w} \in \Delta_G} R(\mathbf{w}) \leq 2 \sup_{\mathbf{w} \in \Delta_G} |R(\mathbf{w}) - \hat{R}(\mathbf{w})|. \quad (\text{A.7})$$

Substituting (A.1) yields (A.2). □

Accuracy and diversity. For intuition, under squared loss and for a fixed weight vector, the error of a weighted mixture admits the exact accuracy–diversity decomposition:

$$\mathbb{E} \left[\left(\sum_{g=1}^G w_g e_g \right)^2 \right] = \sum_{g=1}^G w_g^2 \mathbb{E}[e_g^2] + \sum_{g \neq h} w_g w_h \mathbb{E}[e_g e_h], \quad (\text{A.8})$$

where e_g denotes subgroup- g prediction error. The cross terms are error cross-moments: disagreement helps only when it comes from subgroups whose errors are not perfectly aligned and whose marginal accuracy is not too low.

Scope for the paper’s two applications. Proposition A-4 is stated in scalar form for clarity. For the two losses used in this paper, each per-item discrepancy can be expanded into a finite collection of scalar-coordinate losses applied to affine functions of $\mathbf{w}$: the ranking surrogate contributes $\binom{M_g}{2} \leq \binom{M}{2}$ pair terms for item g , while the discrete WD objective contributes $N_q - 1$ CDF-coordinate terms per item. If the number and magnitude of these coordinates, their coordinatewise Lipschitz constants, and the per-item summed loss are uniformly bounded, the same argument retains the $\sqrt{\log G/Q}$ dependence, with application-specific constants. We therefore use Proposition A-4 as a scalar prototype for the exact application losses.

A.4. Theorem A-6 (Near-Optimal Transfer Under Moderated Domain Shift)

The following theorem provides a sufficient-condition result for Section 2.4.2.

THEOREM A-6 (Near-Optimal Transfer Under Moderated Domain Shift). *Let $\mathcal{D}(\cdot, \cdot)$ be any discrepancy that is convex in its first argument and K -Lipschitz in each argument with respect to the ℓ_1 -norm (e.g., the Wasserstein-1 distance on a metric space of diameter $D_{\max}$, for which $K = D_{\max}$ is a valid choice).*

For every domain $k \in \{1 \text{ (source)}, 2 \text{ (target)}\}$, subgroup $g \in [G]$, and question $q \in \mathcal{Q}$, define the bias vector

$$\beta_g^{(k)}(q) = p_{L,g}^{(k)}(q) - p_H^{(k)}(q) \in \mathbb{R}^d.$$

Here d is the dimension of the response representation, which is the same in both domains.

Assumption 1 (Moderated proportional bias).²⁵ There exist constants $\rho \geq 0$, $\Delta_{\text{bias}} \geq 0$, $B \geq 0$ (here B bounds the bias vectors and is distinct from the output bound B in (A3)) and scalars $\{\lambda_q\}_{q \in \mathcal{Q}}$ such that, for every q, g ,

$$\beta_g^{(2)}(q) = \lambda_q \beta_g^{(1)}(q) + \varepsilon_{gq}, \quad |\lambda_q - 1| \leq \rho, \quad \|\varepsilon_{gq}\|_1 \leq \Delta_{\text{bias}}, \quad \|\beta_g^{(1)}(q)\|_1 \leq B.$$

Assumption 2 (Bounded label shift). There exists $\Delta_{\text{label}} \geq 0$ such that

$$\mathbb{E}_{q \sim D_Q} [\|p_H^{(1)}(q) - p_H^{(2)}(q)\|_1] \leq \Delta_{\text{label}}.$$

Both domains are evaluated under a common question distribution D_Q over items. Define the domain-specific risk

$$R_k(\mathbf{w}) := \mathbb{E}_{q \sim D_Q} \left[\mathcal{D}(p_L^{(k)}(\mathbf{w}, q), p_H^{(k)}(q)) \right], \quad p_L^{(k)}(\mathbf{w}, q) := \sum_{g=1}^G w_g p_{L,g}^{(k)}(q),$$

and let $\mathbf{w}_k^* \in \arg \min_{\mathbf{w} \in \Delta_G} R_k(\mathbf{w})$.

Then, under Assumptions 1–2,

$$R_2(\mathbf{w}_1^*) \leq R_2(\mathbf{w}_2^*) + 2K(\rho B + \Delta_{\text{bias}}) + 4K \Delta_{\text{label}}.$$

This theorem applies directly only to calibration discrepancies that satisfy the stated Lipschitz assumptions, so applying it to a specific loss requires verifying those assumptions in that discrepancy representation. (Only the Lipschitz property is used in the proof; convexity of $\mathcal{D}$ is assumed for consistency with the calibration framework.) When source and target do not share a common item universe, or when the discrepancy representation has not been shown to satisfy the required Lipschitz property, we use the theorem as guidance. The result also concerns exact reuse of the source-risk optimizer, not the population-share adjustment in Equation (3). We therefore assess transfer empirically.

Proof of Theorem A-6.

Step 1: Controlling cross-domain deviations. For any $\mathbf{w} \in \Delta_G$ and q set

$$\delta_\beta(\mathbf{w}, q) := \sum_{g=1}^G w_g (\beta_g^{(2)}(q) - \beta_g^{(1)}(q)), \quad \delta_H(q) := p_H^{(2)}(q) - p_H^{(1)}(q).$$

Assumption 1 gives $\|\delta_\beta(\mathbf{w}, q)\|_1 \leq \rho B + \Delta_{\text{bias}}$, and Assumption 2 yields $\mathbb{E}_q[\|\delta_H(q)\|_1] \leq \Delta_{\text{label}}$.

Because $p_L^{(k)}(\mathbf{w}, q) = p_H^{(k)}(q) + \sum_g w_g \beta_g^{(k)}(q)$,

$$p_L^{(2)}(\mathbf{w}, q) = p_L^{(1)}(\mathbf{w}, q) + \delta_\beta(\mathbf{w}, q) + \delta_H(q).$$

²⁵ Assumption 1 uses the same multiplier λ_q for every subgroup g within a question. Between the two domains, the pattern of subgroup biases can therefore change only by a common scale factor and a bounded residual.

Step 2: Uniform deviation of risks. For fixed q abbreviate $a_2 = p_L^{(2)}(\mathbf{w}, q)$, $b_2 = p_H^{(2)}(q)$, $a_1 = p_L^{(1)}(\mathbf{w}, q)$, $b_1 = p_H^{(1)}(q)$, where the subscript indexes the domain. Then $a_2 - a_1 = \delta_\beta(\mathbf{w}, q) + \delta_H(q)$ and $b_2 - b_1 = \delta_H(q)$. Using K -Lipschitzness in each argument,

$$|\mathcal{D}(a_2, b_2) - \mathcal{D}(a_1, b_1)| \leq K\|a_2 - a_1\|_1 + K\|b_2 - b_1\|_1 \leq K\|\delta_\beta(\mathbf{w}, q)\|_1 + 2K\|\delta_H(q)\|_1.$$

Taking expectations and substituting the bounds above,

$$|R_2(\mathbf{w}) - R_1(\mathbf{w})| \leq K(\rho B + \Delta_{\text{bias}}) + 2K\Delta_{\text{label}} \quad \forall \mathbf{w} \in \Delta_G. \quad (\text{A.9})$$

Step 3: Comparing the two optimizers. Let $\Delta(\mathbf{w}) := R_2(\mathbf{w}) - R_1(\mathbf{w})$. Because $\mathbf{w}_1^*$ minimizes R_1 ,

$$\begin{aligned} R_2(\mathbf{w}_1^*) &= R_1(\mathbf{w}_1^*) + \Delta(\mathbf{w}_1^*) \leq R_1(\mathbf{w}_2^*) + \Delta(\mathbf{w}_1^*) \\ &= R_2(\mathbf{w}_2^*) - \Delta(\mathbf{w}_2^*) + \Delta(\mathbf{w}_1^*) \leq R_2(\mathbf{w}_2^*) + |\Delta(\mathbf{w}_1^*)| + |\Delta(\mathbf{w}_2^*)|. \end{aligned}$$

Applying the bound in Equation (A.9) to each $\Delta(\mathbf{w}_j^*)$, $j = 1, 2$,

$$|\Delta(\mathbf{w}_j^*)| \leq K(\rho B + \Delta_{\text{bias}}) + 2K\Delta_{\text{label}},$$

we obtain

$$R_2(\mathbf{w}_1^*) \leq R_2(\mathbf{w}_2^*) + 2K(\rho B + \Delta_{\text{bias}}) + 4K\Delta_{\text{label}}.$$

□

A.5. Proposition A-7 (Monotonic Improvement Under Uniform Upgrades)

PROPOSITION A-7 (Monotonic Improvement Under Uniform Upgrades). Let $\{\mathcal{Z}_q^L\}_{q=1}^Q$ be the original LLM outputs and $f(\mathbf{w}) = \sum_{q=1}^Q \mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H)$. Let $\{\tilde{\mathcal{Z}}_q^L\}_{q=1}^Q$ be new LLM outputs that satisfy $\mathcal{D}(\tilde{\mathcal{Z}}_q^L(\mathbf{w}), \mathcal{Z}_q^H) \leq \mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H)$, $\forall q, \forall \mathbf{w} \in \Delta_G$. Denote by $\tilde{f}(\mathbf{w})$ the corresponding calibration objective. Then $\inf_{\mathbf{w} \in \Delta_G} \tilde{f}(\mathbf{w}) \leq \inf_{\mathbf{w} \in \Delta_G} f(\mathbf{w})$.

Proposition A-7 formalizes a conditional sense in which HuMCal is compatible with model upgrades. We call the new model outputs an improvement if they are uniformly closer to the human benchmark under the same discrepancy $\mathcal{D}$ used by HuMCal:

$$\mathcal{D}(\tilde{\mathcal{Z}}_q^L(\mathbf{w}), \mathcal{Z}_q^H) \leq \mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H) \quad \forall q, \forall \mathbf{w} \in \Delta_G.$$

Under this condition, any such improvement in the underlying LLM outputs weakly reduces the optimal value of the calibration objective, so the best achievable fit after calibration cannot worsen. The condition applies to every item and feasible mixture under $\mathcal{D}$; improvement at the individual checkpoints alone is insufficient, and the result does not imply improvement in AR. This is a sufficient condition and need not hold for arbitrary model updates: if a future model improves some items but degrades others under $\mathcal{D}$, the calibrated optimum can move in either direction.

Proof of Proposition A-7. Let

$$f(\mathbf{w}) = \sum_{q=1}^Q \mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H), \quad \tilde{f}(\mathbf{w}) = \sum_{q=1}^Q \mathcal{D}(\tilde{\mathcal{Z}}_q^L(\mathbf{w}), \mathcal{Z}_q^H).$$

By the hypothesis of the proposition,

$$\mathcal{D}(\tilde{\mathcal{Z}}_q^L(\mathbf{w}), \mathcal{Z}_q^H) \leq \mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H), \quad \forall q = 1, \dots, Q, \forall \mathbf{w} \in \Delta_G.$$

Summing over q yields

$$\tilde{f}(\mathbf{w}) \leq f(\mathbf{w}), \quad \forall \mathbf{w} \in \Delta_G.$$

Taking infima over $\mathbf{w} \in \Delta_G$ on both sides gives

$$\inf_{\mathbf{w} \in \Delta_G} \tilde{f}(\mathbf{w}) \leq \inf_{\mathbf{w} \in \Delta_G} f(\mathbf{w}).$$

□

The guarantee concerns the calibration sample, so out-of-sample behavior still requires the checks in Appendix C.

A.6. Finite-Step Softmax Implementation

The results above concern an exact minimizer $\mathbf{w}^*$ of the closed-simplex program in Equation (HuMCal-OPT). Our implementation approximates it: we set $\mathbf{w} = \text{softmax}(\boldsymbol{\theta})$ and fit $\boldsymbol{\theta}$ with finite-step Adam or AdamW, so the returned weights are an interior point of the simplex and carry no exact global-optimality or KKT certificate. Writing the unregularized objective as $f(\mathbf{w})$, softmax ranges over the relative interior of the simplex, so at $\lambda = 0$,

$$\inf_{\boldsymbol{\theta} \in \mathbb{R}^G} f(\text{softmax}(\boldsymbol{\theta})) = \min_{\mathbf{w} \in \Delta_G} f(\mathbf{w}),$$

but a boundary minimum need not be attained at finite logits. All main analyses use this finite-step fit at $\lambda = 0$; the comparisons with uniform averaging and the best single checkpoint are therefore comparisons against it.

In all three empirical setups, the implemented softmax fit attains a lower mean calibration loss per retained pair than every fixed-weight comparator: direct elicitation, post-stratification, and the best single persona in the two demographic setups, and uniform averaging and the best single checkpoint in the five-checkpoint setup. In the five-checkpoint setup, the exact closed-simplex solution assigns positive weight to only two checkpoints and matches the softmax fit’s mean calibration loss to four decimal places, but attains a lower AR: 0.669, against 0.744 for the softmax fit and 0.676 for uniform averaging. Retaining small positive weights through softmax enlarges the support of the calibrated mixture, from two checkpoints to five, and with it the share of human-listed pairs the combination covers. Appendix D.5 gives that coverage–ordering decomposition. On the pairs the exact solution also covers, the two rules order identically, and the gain in AR comes entirely from the additional pairs that only the low-weight checkpoints reach.

Appendix B: Intuition Through Hypothetical Numerical and Geometric Illustrations

B.1. Hypothetical Numerical Illustration of the Mechanism

We use a toy example to show how calibration can correct counteracting subgroup errors. Consider estimating Dallas-metro consumers’ willingness-to-pay (WTP) for Coca-Cola. The team has exactly one costly human benchmark, the \$5.00 population-level survey figure from last quarter, and no information for individual subgroups. The immediate goal is to predict WTP for Coca-Cola in other metropolitan areas (e.g., Los Angeles and New York) as well as for other soft drinks such as Pepsi, Sprite, and Dr Pepper.

An LLM is queried once for Dallas as a whole and again for two virtual subgroups, “Dallas female” and “Dallas male.” For illustration, suppose the true and unobserved WTP is \$5.00 for women and \$5.00 for men, with the population split exactly 50:50, so the metro’s observed WTP is also \$5.00. The model, however, replies \$4.50 for the aggregate, \$4.00 for women, and \$8.00 for men. A naive 50:50 demographic average therefore yields \$6.00, overshooting the benchmark, whereas the direct aggregate prompt undershoots.

HuMCal learns calibration weights from a single aggregate benchmark. In this toy example it assigns weight 0.75 to the lower estimate provided by women and 0.25 to the higher estimate provided by men, so that $0.75 \times \$4.00 + 0.25 \times \$8.00 = \$5.00$, exactly reproducing the human figure. With two candidates these weights are the only solution. If the aggregate reply were also included as a candidate, the two calibration equations would no longer pin down three weights, and several weightings would reproduce the benchmark. This weighted average matches the aggregate benchmark without requiring subgroup-level human data. The same logic readily extends to additional or more granular subgroups.

The method is purposefully centered on the analyst-defined aggregate because strategic decisions (e.g., pricing, media spend, policy targets) are made at that level, and collecting exhaustive subgroup labels can be prohibitively expensive. If subgroup-specific human data (female and male subgroups in the Dallas metro) later appear, HuMCal can be applied within each subgroup (Appendix D.6).

This example can also illustrate why weights may transfer. If subgroup errors relative to the human benchmark data change little across settings (here, women understating and men overstating), source-learned weights can remain useful. Theorem A-6 gives one sufficient condition for this type of transfer for discrepancies satisfying the assumptions in Appendix A.4. Our empirical analyses in Sections 3 and 4 are consistent with this intuition, but they do not by themselves verify those assumptions.

B.2. Geometric Explanation of HuMCal

The simplex constraint in Equation (HuMCal-OPT) implies that the aggregate LLM output is a convex combination of subgroup outputs. Fix a question q . Each subgroup output $\mathcal{Z}_{q,g}^L$ can be viewed as a point in the output space for that question (e.g., a scalar prediction, a vector of category probabilities/CDF values, or a vector of pairwise ranking margins). As $\mathbf{w}$ varies over the simplex $\Delta_G := \{\mathbf{w} \in \mathbb{R}_+^G : \sum_{g=1}^G w_g = 1\}$, the feasible aggregates $\mathcal{Z}_q^L(\mathbf{w}) = \sum_{g=1}^G w_g \mathcal{Z}_{q,g}^L$ trace out the convex hull of these subgroup points. The calibration selects $\mathbf{w}^*$ to minimize the discrepancy $\mathcal{D}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H)$ summed across q . When $\mathcal{D}$ is convex in its first argument, the objective is convex over Δ_G , so any local optimum is global (Theorem A-1); strict convexity is one sufficient route to uniqueness, but uniqueness can also occur without strict convexity.

Figure B.1 illustrates the geometry for two questions and three subgroups. In panel (a), the subgroup outputs for each question (colored marks) form a triangle; any point inside the triangle is a convex mixture of those outputs. The blue star is the human benchmark. The mixture induced by $\mathbf{w}^*$ is shown as a pie chart whose wedges are the calibrated weights. Because the calibration uses a single weight vector across questions, $\mathbf{w}^*$ generally represents a compromise that minimizes total loss across questions; it need not be the closest point to the benchmark for each question considered in isolation.

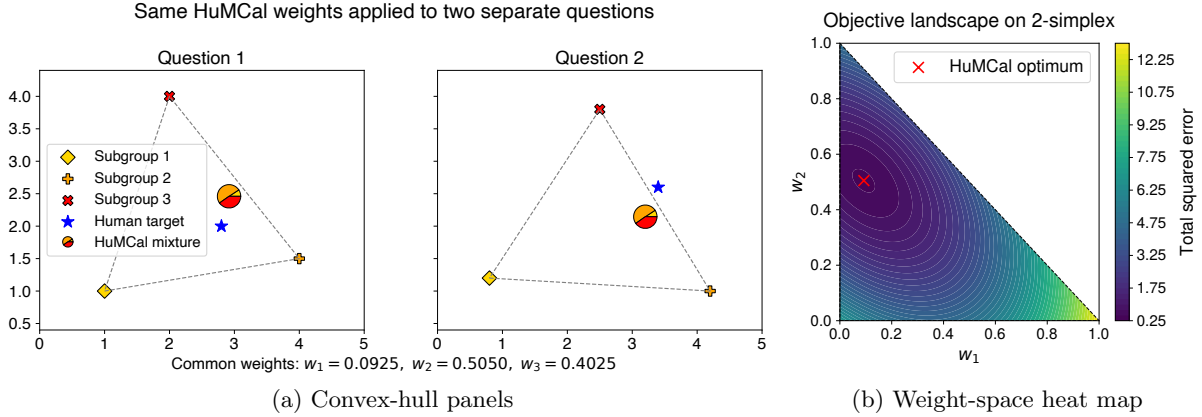


Figure B.1 Geometric explanation of HuMCal's simplex calibration. In (a), each question's subgroup outputs form a convex hull, and a pie chart marks the mixture induced by the calibrated weights. In (b), darker shading indicates lower total error.

Panel (b) shows the same problem in weight space (the simplex). Each point corresponds to a feasible $\mathbf{w} \in \Delta_3$, and shading represents the total calibration loss over both questions. Convexity implies that every local minimum is global, although the minimizer need not be unique.

Connection to baselines. Common heuristics are distinct fixed points on the simplex: uniform averaging sits at the interior centroid; direct elicitation is the aggregate-prompt vertex; and post-stratification or a single chosen subgroup are other fixed points. The exact closed-simplex HuMCal program minimizes the calibration objective over that same simplex, so its calibration loss is no greater than at any of those fixed points (Proposition A-2; Corollary A-3).

Appendix C: Out-of-Sample Sensitivity Analyses and Subgroup-Design Diagnostics

The main-text transfer experiments (Steps 6–7) show that HuMCal generally improves upon fixed-weight baselines out-of-sample. A natural follow-up is: how sensitive are these gains to subgroup count and composition, regularization choices, and model upgrades? This appendix addresses these questions through a series of diagnostics.

Theoretical motivation. Two forces determine whether calibrated weights remain useful beyond the source sample. Finite calibration data limit how many candidate weights can be estimated reliably, and domain shift can change the relative accuracy of those candidates. Proposition A-4 characterizes the first force through a $\sqrt{\log G/Q}$ estimation term. Theorem A-6 gives sufficient conditions for the second. Because the transfer conditions require target-domain information, we evaluate both forces empirically.

Roadmap. Appendix C.1 records a simple decomposition of target-domain relative performance into a source-domain difference and a differential shift term (Equation (C.1)). Appendices C.2–C.3 probe finite-sample generalization via regularization sensitivity (Figure C.1), subgroup-count learning curves under random addition (Figure C.2), and a randomly-shuffled-subgroup robustness check (Figure C.3). Appendix C.4 summarizes composition effects through a pooled regression diagnostic relating marginal out-of-sample gains to subgroup accuracy, diversity, and standalone agreement shifts (Table C.1). Appendix C.5 provides weight diagnostics (Table C.2; Figure C.4). Appendix C.6 discusses practical subgroup-design guidance, and Appendix C.7 reports a checkpoint-upgrade evaluation (Figure C.5).

Diagnostic settings. The subgroup-count and regression analyses use the 13-candidate micro-market setting (12 personas plus direct elicitation), which allows us to vary subgroup count while keeping brand coverage broadly comparable. The regularization check uses the Dallas-metro persona setting. The no-signal and checkpoint-upgrade checks use the five-checkpoint setting. Weight diagnostics cover all three settings.

C.1. A Simple Decomposition of Out-of-Sample Transfer Performance

To connect the empirical diagnostics below to the transfer discussion in Section 2.2.4, we use an algebraic identity that allows both the calibrated rule and its comparator to use different weight vectors in the source and target domains. Let $R_S(\mathbf{w})$ and $R_T(\mathbf{w})$ denote population discrepancy (under $\mathcal{D}$) in source domain S and target domain T . Let $\hat{\mathbf{w}}_S$ be the fitted source weighting and $\hat{\mathbf{w}}_{S \rightarrow T}$ the weighting actually applied in the target. Let $\mathbf{b}_S$ and $\mathbf{b}_T$ denote the baseline weightings actually applied in the two domains. Under direct reuse, $\hat{\mathbf{w}}_{S \rightarrow T} = \hat{\mathbf{w}}_S$; Equation (3) can instead produce a target-adjusted calibrated vector. Likewise, $\mathbf{b}_T = \mathbf{b}_S$ for a fixed baseline but can differ under target-specific post-stratification. Then

$$R_T(\hat{\mathbf{w}}_{S \rightarrow T}) - R_T(\mathbf{b}_T) = \underbrace{[R_S(\hat{\mathbf{w}}_S) - R_S(\mathbf{b}_S)]}_{\text{source-domain difference}} + \underbrace{[\Delta_{\text{shift}}(\hat{\mathbf{w}}_S, \hat{\mathbf{w}}_{S \rightarrow T}) - \Delta_{\text{shift}}(\mathbf{b}_S, \mathbf{b}_T)]}_{\text{differential shift term}}, \quad (\text{C.1})$$

$$\Delta_{\text{shift}}(\mathbf{w}_S, \mathbf{w}_T) := R_T(\mathbf{w}_T) - R_S(\mathbf{w}_S).$$

The first term is negative when calibration lowers source-domain population risk relative to the source baseline, but it can be positive when the empirically fitted weights overfit the source calibration split. The second is a differential shift term: it compares how the calibrated and baseline procedures change from source to target, incorporating both domain shift and any source-to-target change in each procedure’s weight vector.

Under direct reuse the two arguments coincide, $\Delta_{\text{shift}}(\mathbf{w}, \mathbf{w}) = R_T(\mathbf{w}) - R_S(\mathbf{w})$; carrying both covers the adjustment in Equation (3) and target-specific post-stratification.

This identity motivates the out-of-sample sensitivity analyses in this appendix: learning curves and composition diagnostics probe diminishing returns and finite-sample stability, while transfer experiments probe whether subgroup credibility remains stable enough for a favorable source-domain difference to persist in the target domain.

C.2. Explicit Regularization: Limited Gains beyond the Simplex Constraint

Figure C.1 reports how HuMCal’s performance in the market-structure application changes as we vary the strength of the logit-space ℓ_2 penalty. Because Equation (HuMCal-OPT) already restricts weights to a simplex (a convex mixture), it benefits from implicit regularization by staying within the convex hull of subgroup predictions.

As an auxiliary sensitivity check, we add a squared ℓ_2 penalty to the logits. For $\lambda > 0$, we define the exact penalized target as

$$\boldsymbol{\theta}_\lambda^* \in \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^G} \frac{1}{|\mathcal{P}|} \sum_{(q,j,k) \in \mathcal{P}} \ell_{q,jk}(\text{softmax}(\boldsymbol{\theta})) + \lambda \|\boldsymbol{\theta}\|_2^2, \quad \mathbf{w}_\lambda^* = \text{softmax}(\boldsymbol{\theta}_\lambda^*), \quad (\text{C.2})$$

where $\ell_{q,jk}$ is the pairwise logistic loss and $\mathcal{P}$ is the set of retained pairs. We approximate this objective with finite-step optimization. The logit-parameterized objective need not be convex in $\boldsymbol{\theta}$, so the finite-step solution is not guaranteed to be a global minimizer. The $\lambda > 0$ fits are a separate penalized sensitivity specification and do not inherit the exact program’s baseline-safety guarantee. All main analyses use the unregularized specification, $\lambda = 0$, and no positive grid value improves out-of-sample agreement consistently across the displayed domains. At every positive grid value the fitted weights coincide with uniform averaging over the 85 candidates (largest weight 0.0118 against 1/85), so the positive- λ points compare the unregularized fit with uniform weighting.

C.3. Finite-Sample Robustness: Diminishing Returns and No-Signal Candidates

We next group diagnostics that probe finite-sample generalization risk (overfitting) within a fixed calibration domain: sensitivity to subgroup count, robustness to adding no-signal candidate subgroups, and sensitivity to explicit regularization (Appendix C.2).

C.3.1. Subgroup-count learning curves. Our candidate pool contains 12 gender-by-age personas and direct elicitation. For each $n = 1, \dots, 13$, the diagnostic uses 100 retained candidate sets of size n . Candidates are sampled without replacement within each set, and sets may recur across draws. A set is retained only if its selected personas have positive total population share in both the calibration and evaluation regions. For each retained set, we refit HuMCal on the fixed calibration split and evaluate the fitted weights in the corresponding transfer setting. For the post-stratification baseline, direct elicitation receives zero population weight, and evaluation-region population shares are renormalized over the selected personas. In the geographic panel, these are the target-region shares; in the temporal and industry panels, the calibration and evaluation geography is the same. The curves plot mean improvements in AR over direct elicitation across the 100 retained draws at each n . Under Proposition A-4’s idealized same-domain assumptions and fixed envelopes,

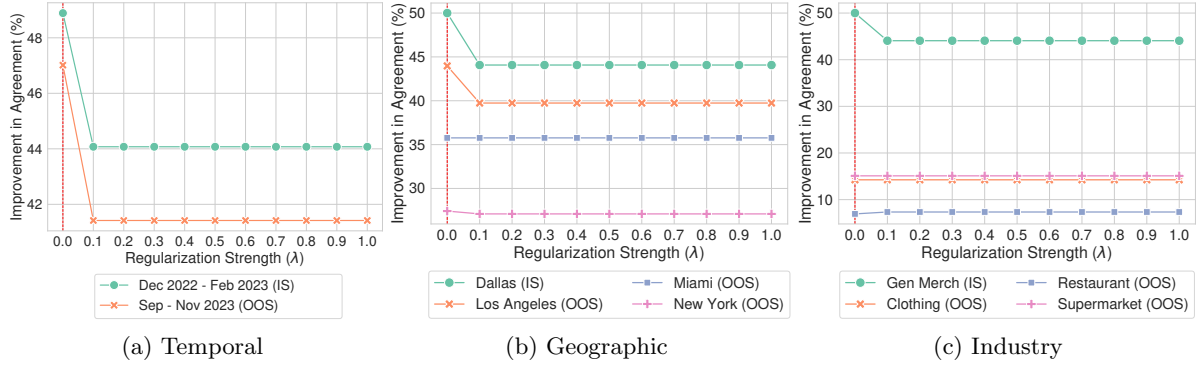


Figure C.1 Regularization sensitivity for HuMCal in the Dallas-metro persona setting. We vary the strength $\lambda \in \{0, 0.1, \dots, 1.0\}$ of an ℓ_2 penalty on the pre-softmax logits that parameterize the simplex weights. The plotted quantity is improvement in agreement over direct elicitation, in-sample (green) and out-of-sample (orange/blue/pink) across temporal, geographic, and industry transfers. The red dashed line marks the unregularized specification $\lambda = 0$ used in all main analyses. No positive value on the grid improves out-of-sample agreement consistently across domains; where it changes performance, positive λ lowers or leaves agreement unchanged in most targets and yields only a small improvement for restaurants, consistent with shrinkage toward uniform weights.

the worst-case loss-generalization term grows as $\sqrt{\log G/Q}$. The transfer-AR curves below are an empirical diagnostic. Improvements in AR rise rapidly with subgroup count in every panel (Figure C.2). The temporal and clothing curves flatten after about eight candidates, whereas the restaurant and supermarket curves continue to rise more slowly through $n = 13$. The geographic curve is negative for the smallest candidate sets, peaks at eight candidates, and declines overall thereafter.

C.3.2. Robustness to uninformative subgroups. To test whether HuMCal down-weights uninformative candidates, we augment the five-LLM checkpoint setting with five synthetic subgroups whose brand rankings carry no systematic relationship to the human benchmark data. To rule out shorter lists or reduced brand coverage as an alternative explanation for their low weight, we construct each synthetic subgroup by shuffling only the ranking order within one real checkpoint’s own focal-brand lists. Each synthetic subgroup therefore preserves the checkpoint’s exact brand set, and thus its pairwise coverage, but replaces the original ordering with a random one. Pairing the five checkpoints with five independently shuffled lists yields five independent random subgroups. We then calibrate HuMCal on the augmented pool and repeat the entire construction 50 times.

Figure C.3 shows that the random subgroups consistently receive near-zero weight across draws, while the calibrated mass remains concentrated on the true LLM checkpoints. Because coverage is matched by construction, the synthetic subgroups’ low weights cannot be explained by shorter lists or narrower brand coverage. What the experiment establishes is that they never enter the top-weighted set: the calibrated mass concentrates on two genuine checkpoints, and every remaining candidate—synthetic and genuine alike—receives near-zero weight.

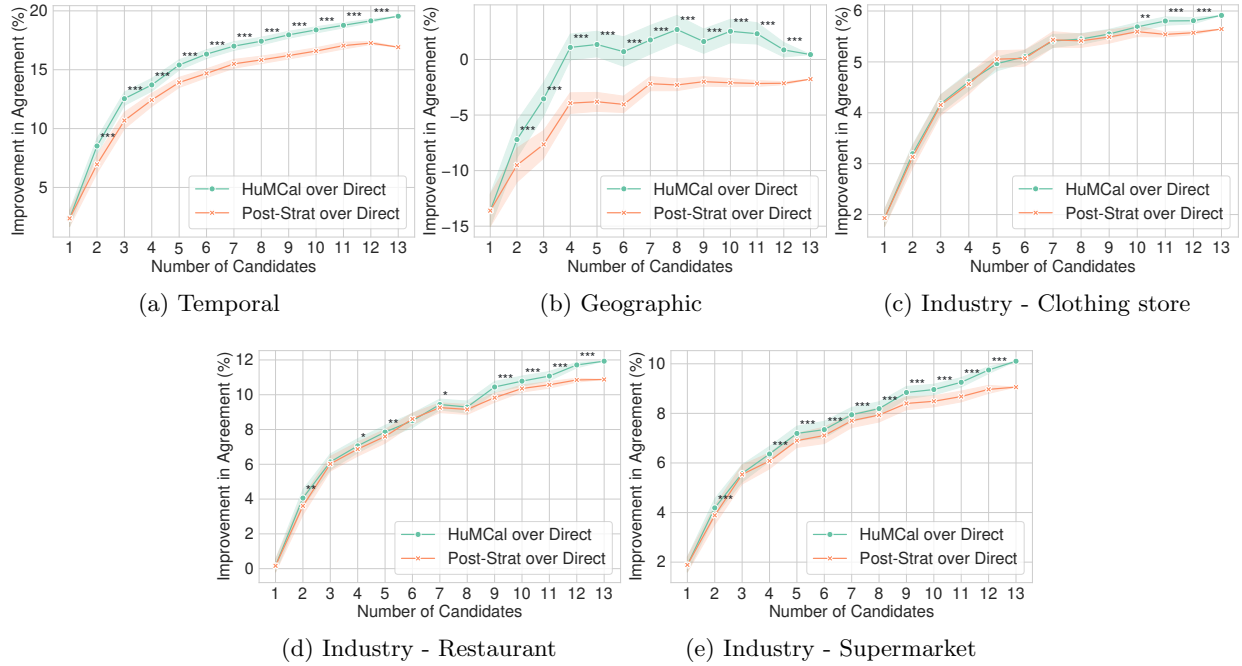


Figure C.2 Random candidate-set learning curves in the Park Cities (Dallas)–\$60K–\$75K setting. Each panel corresponds to a different transfer domain (temporal, geographic, industry). The y-axis reports improvements in AR over direct elicitation as a function of the number of included candidates n . Curves average over 100 retained candidate sets at each n . Stars summarize one-sided Monte Carlo t -tests across the 100 retained candidate sets at each n of whether HuMCal’s improvement over direct exceeds post-stratification’s improvement over direct ($p < 0.1^*$, $p < 0.05^{**}$, $p < 0.01^{***}$). We omit significance tests at $n = 1$ and $n = 13$ because the difference between HuMCal and post-stratification does not vary across draws at those endpoints. Shaded bands are bootstrap 95% confidence intervals for the mean across the 100 retained candidate sets at each n .

Note. These statistics describe variation from randomized candidate-set selection conditional on the observed dataset; they are not population-sampling tests.

C.4. Regression Diagnostic: How Subgroup Characteristics Relate to Marginal Gains

To complement the curve-based evidence above, we report a pooled regression diagnostic that summarizes how subgroup accuracy, ranking diversity, and the standalone agreement shift relate to marginal improvements beyond subgroup count. Subgroup accuracy and ranking diversity use source data; the standalone agreement shift also uses target benchmark data and describes observed transfer outcomes.

For this analysis, we focus on the Park Cities (Dallas)–\$60K–\$75K setting with $G = 13$ candidate subgroups (12 demographic-persona subgroups and one direct-elicitation subgroup). For each repetition $r = 1, \dots, 50$ and each subgroup-count index $n \in \{2, \dots, 12\}$, we first sample a base set $\mathcal{S}_{n-1}$ of size $n - 1$ uniformly at random from the G candidates. We then select one additional subgroup g_n from the remaining candidates using one of three prespecified rules: random, most similar, or most different, where dissimilarity is measured by Kendall- τ distances between subgroup rankings computed on the in-sample calibration questions. For the most-similar and most-different rules, a candidate’s dissimilarity to $\mathcal{S}_{n-1}$ is its mean Kendall- τ distance to the members of that set; the regressor defined below uses the minimum distance instead. We refit HuMCal

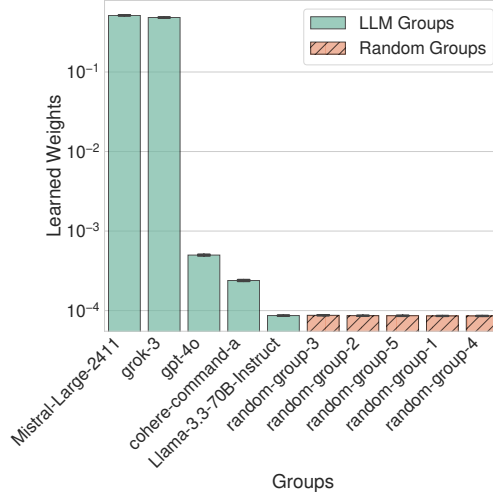


Figure C.3 Robustness to uninformative subgroups. Each synthetic subgroup shuffles the ranking order within one real checkpoint’s own brand lists, matching that checkpoint’s pairwise coverage exactly. Adding the five synthetic subgroups to the five-checkpoint pool yields near-zero learned weights on them. Bars are means over 50 draws and error bars are bootstrap 95% confidence intervals for the mean across those draws.

on the fixed in-sample calibration split using (i) $\mathcal{S}_{n-1}$ and (ii) $\mathcal{S}_n = \mathcal{S}_{n-1} \cup \{g_n\}$, and compute the marginal gain in agreement:

$$\Delta \text{AR}_n := \text{AR}(\mathcal{S}_n) - \text{AR}(\mathcal{S}_{n-1}).$$

We evaluate ΔAR_n both in-sample (three source domains: industry/general merchandise, spatial/Dallas, and temporal/December 2022–February 2023) and out-of-sample (five transfer domains: three industry targets, spatial/East Collin County, and temporal/September–November 2023). This yields 4,950 in-sample regression rows ($11 \times 3 \times 3 \times 50$) and 8,250 out-of-sample regression rows ($11 \times 5 \times 3 \times 50$). The marginal gain enters the regression below in the same units as the reported improvement measures: as a percentage of the remaining room for improvement over direct elicitation, $100 \Delta \text{AR}_n / (1 - \text{AR}_0)$, with AR_0 denoting AR under direct elicitation.

For each regression row (each added subgroup g_n), we define: (i) *subgroup AR*, $\text{AR}(g_n)$, as the standalone agreement of subgroup g_n on the source calibration split; (ii) *minimum ranking distance* $d_{\min}(g_n) = \min_{g \in \mathcal{S}_{n-1}} d_{\tau}(g_n, g)$ as the minimum Kendall- τ distance between g_n and the previously selected set; and (iii) for the out-of-sample regression only, the *standalone agreement shift* $\Delta \widetilde{\text{AR}}_{T-S}(g_n) = \widetilde{\text{AR}}_T(g_n) - \widetilde{\text{AR}}_S(g_n)$, where $\widetilde{\text{AR}}$ denotes min-max normalized subgroup agreement within each domain. Positive values indicate that the subgroup’s relative standalone agreement is higher in the target domain, while negative values indicate that it is lower. We standardize all regressors to have zero mean and unit variance. Both specifications include $\text{target} \times n \times \text{selection-strategy}$ fixed effects.

Table C.1 reports uncertainty across 10,000 repetitions of the candidate-set selection experiment, preserving shared draws across evaluations. In Column (1), the industry and spatial exercises share the same source calibration split and the same candidate draws, so their rows coincide. Counting this shared block once leaves 3,300 rows; the reported regression includes the general-merchandise source split twice. Table C.1

Table C.1 Descriptive regression diagnostic (main effects): correlates of the normalized marginal gain from adding one subgroup and refitting HuMCal. Column (1) uses in-source evaluations; Column (2) uses out-of-sample transfer evaluations and additionally includes the standalone agreement shift.

	Dependent Variable: $100 \Delta \text{AR}_n / (1 - \text{AR}_0)$	
	(1) In-Sample	(2) Out-of-Sample
Accuracy (Subgroup AR)	0.972*** (0.050)	0.800*** (0.028)
Diversity (Minimum Ranking Distance)	0.957*** (0.097)	0.013 (0.047)
Standalone Agreement Shift	–	1.343*** (0.036)
Constant	0.667*** (0.201)	0.757*** (0.192)
Fixed Effects	target $\times$ n $\times$ selection-strategy	
Regression Rows	4,950	8,250
R-squared	0.595	0.419
Adj. R-squared	0.587	0.407

Notes: Standard errors (parentheses) are the standard deviation of coefficients over 10,000 complete candidate-set selection experiments, preserving shared draws across evaluations. Stars denote two-sided normal-approximation tests of a zero expected coefficient under this selection procedure, conditional on the fixed human and LLM records and fitted subset scores: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. They do not support population or causal inference. Both specifications include target $\times$ n $\times$ selection-strategy fixed effects. In Column (1), the industry and spatial in-sample rows coincide; counting their shared block once leaves 3,300 of the 4,950 regression rows. The standalone agreement shift uses target benchmark outcomes and is unavailable before they are observed.

shows that in-sample marginal gains are positively associated with subgroup accuracy and ranking diversity (larger minimum Kendall- τ distance). Out-of-sample marginal gains are positively associated with subgroup accuracy and the standalone agreement shift. The estimated out-of-sample coefficient on minimum ranking distance is 0.013, near zero relative to the other estimates.

C.5. Weight Diagnostics

We next provide two descriptive diagnostics of the fitted weights: (i) weights track standalone subgroup ordering agreement among covered pairs, and (ii) randomizing the rankings of high-weight subgroups degrades agreement much more than randomizing those of low-weight subgroups. These weights measure how much each subgroup helps reconstruct the aggregate human benchmark. A subgroup can carry weight because it tracks the benchmark closely on its own, because it covers pairs that other subgroups leave uncovered, or because its errors offset theirs.

C.5.1. Weights track standalone subgroup covered-pair agreement. For each subgroup g we compute (i) its calibrated weight w_g and (ii) its standalone covered-pair agreement c_g . The latter is the number of concordantly ordered human-listed pairs divided by the number of concordant, discordant, and tied pairs for which the subgroup supplies both brands; tied pairs remain in the denominator as nonconcordant. Unlike the headline AR, c_g excludes human-listed pairs that the subgroup does not cover. This diagnostic therefore isolates ordering quality conditional on coverage. Table C.2 reports Spearman rank correlations between w_g and c_g . Positive associations indicate that subgroups with higher covered-pair agreement receive higher weight.

Table C.2 shows that calibrated weights are positively associated with standalone subgroup covered-pair agreement in all three settings. The association is statistically distinguishable from zero in the 85-subgroup and five-checkpoint settings. In the smaller Park Cities (Dallas)–\$60K–\$75K micro-market (13 subgroups), the point estimate remains positive ($\rho = 0.385$) but is not distinguishable from zero at conventional levels ($p = 0.196$). Because both measures use the same calibration benchmark, this association is descriptive.

Table C.2 Weight–covered-pair agreement association: Spearman correlation between calibrated weights (w_g) and standalone subgroup covered-pair agreement (c_g) in the calibration domain. Positive correlations indicate that HuMCal assigns higher weight to subgroups with higher standalone agreement conditional on pair coverage. For the two smaller settings, the reported p -values come from two-sided permutation tests over rank assignments: exact over all $5!$ assignments for the five checkpoints, and Monte Carlo with 10^7 draws for the 13 subgroups. The

85-subgroup row uses the standard asymptotic Spearman test.

Setting	Spearman ρ	p -value
Dallas metro—85 subgroups (income $\times$ gender $\times$ age + direct)	0.5798	< 0.001
Park Cities (Dallas)–\$60K–\$75K—13 subgroups (gender $\times$ age + direct)	0.3846	0.196
Dallas metro—5 LLM models	1.0000	0.017

C.5.2. High-weight subgroups matter most. Starting from the fully calibrated model, we probe how much the aggregate depends on each subgroup’s ranking content by randomizing that subgroup’s ranking while holding its calibrated weight fixed. We then re-evaluate the agreement, and average over 100 repeats. We report the loss as a share of the calibrated aggregate’s agreement: the calibrated agreement minus the agreement after randomization, divided by the calibrated agreement. We perturb subgroups in descending order of calibrated weight.

Figure C.4 shows that randomizing the highest-weighted subgroups lowers agreement the most, while randomizing lower-weighted subgroups has little effect. Because a perturbed subgroup keeps its weight but contributes uninformative pairwise votes, this diagnostic measures how sensitive the aggregate is to randomized ranking content.²⁶

We examined the following three empirical contexts:

- **Dallas metro, 85 subgroups** (84 demographic personas, income $\times$ gender $\times$ age, plus direct elicitation): Given the high number of subgroups, perturbations occur in groups of three. Randomizing the top three subgroups decreases agreement by approximately 13.1% of its calibrated level, while perturbing subsequent groups has a diminishing impact (0.9% for groups 4–6 and negligible thereafter).
- **Park Cities (Dallas)–\$60K–\$75K, 13 subgroups** (12 demographic personas plus direct elicitation, finer income tier within a micro-market): Randomizing the two highest-weighted subgroups reduces agreement by approximately 24.1%, with subsequent perturbations resulting in sharply diminishing reductions (below 2% in magnitude and eventually negligible).

²⁶ A complementary check that instead omits each subgroup and renormalizes the remaining calibrated weights changes agreement only modestly in these settings. This result suggests that the calibrated aggregate does not hinge on any single subgroup.

- **Dallas metro, 5 checkpoints** (representing five model families or checkpoints): Randomizing the ranking of the top-weighted model (**Mistral-Large-2411**) reduces AR by about 27.3%, followed by smaller effects from **Grok-3** (approximately 7.1%) and **GPT-4o** (approximately 1.6%); **Cohere-command-a** and **Llama-3.3-70B-Instruct** have negligible effects (about 1% or less in magnitude).

This analysis shows that calibrated weights help identify the subgroup outputs that most influence the fitted aggregate. Higher weights are associated with better ordering agreement on covered pairs, and randomizing the highest-weighted subgroups’ rankings produces the largest losses in aggregate agreement.

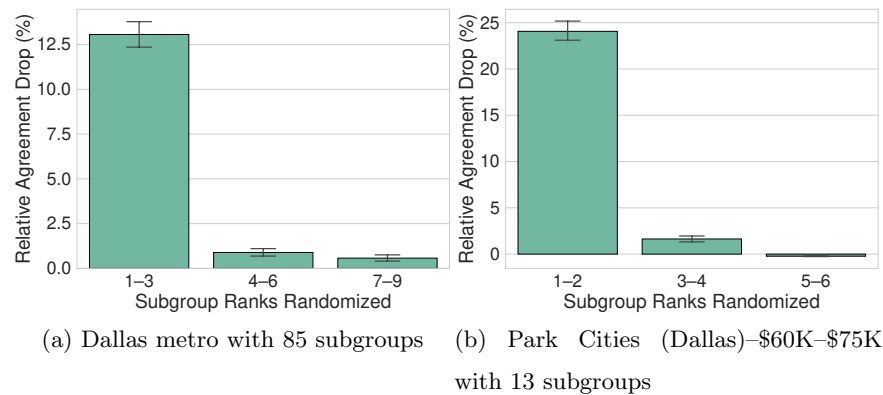
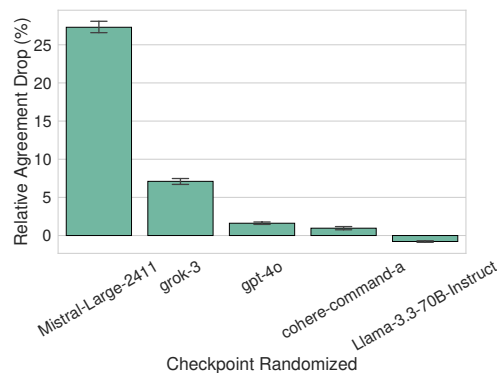


Figure C.4 Ranking-randomization sensitivity: change in agreement, as a percentage of the calibrated aggregate’s agreement, when a subgroup’s ranking is randomized while its calibrated weight is held fixed and the subgroup is retained, taken in descending calibrated-weight order and averaged over 100 repeats. Randomizing the highest-weight subgroups causes the largest agreement losses, while randomizing low-weight subgroups has little effect. Each bar randomizes only the subgroups in that rank block, always starting from the fully calibrated aggregate; the negative bars in panels (b) and (c) reflect small improvements. Error bars are bootstrap 95% confidence intervals for the mean across the 100 repeats.

C.6. Practical Subgroup Design: Open-Ended vs. Menu-Based Tasks

The learning curves and robustness checks above show that adding subgroups can deliver large early gains but that returns usually diminish once the most informative subgroups are included.

Fine demographic personas can surface information that a coarse prompt overlooks, but they can also inject noise if the LLM was trained on few examples from that niche. Whether finer personas pay off depends mainly on how large the set of possible answers is. We therefore distinguish two broad task families.

- (1) *Open-ended tasks (large answer set)*. Questions such as “List brands most often co-purchased with Target” or “Suggest competitors for a new soda” span thousands of possible tokens. Fine demographic personas can provide complementary retrieval cues. For example, they may surface brands that a coarse prompt omits. HuMCal uses the calibration data to determine how much weight these additional outputs receive.
- (2) *Menu-based tasks*. For multiple-choice or Likert items, the full option set is already visible to the model. Here, bias rather than omission dominates, and extra personas may inflate variance faster than they shrink bias. Stromberg et al. (2025) document accuracy losses when naïvely averaging finer demographic groups on such fixed-menu questions.

We offer this contrast as a design hypothesis, supported by one open-ended and one menu-based application. Testing it broadly would require several tasks of each type, which we leave for future work.

To capture the benefits of finer subgrouping while limiting its added variance, we recommend two design rules:

- (1) *Model averaging*. If possible, elicit the same responses from multiple independent LLMs. Averaging across checkpoints lowers model-specific idiosyncrasies before calibration (Santurkar et al. 2023, Schoenegger et al. 2024).
- (2) *Aggregate-as-subgroup guarantee*. Include direct aggregate elicitation as a candidate in the subgroup list. Appendix A.2 shows that, once this cell is included in the subgroup list, direct elicitation becomes a feasible simplex point, so the exact optimizer’s calibration loss cannot be worse while it remains free to exploit complementary signal in finer cells. This statement does not imply dominance in external metrics such as AR or normalized agreement, nor out-of-sample superiority under distribution shift.

In practice, analysts often lack prior knowledge of which cells are truly informative. For open-ended tasks, begin with a broad but substantively motivated candidate set (i.e., demographic, methodological, or both), and let the optimizer allocate weight according to agreement with the calibration data. For menu-based tasks, where extra variants tend to add variance, use a smaller set of reliable variants. In our applications, this perspective helps interpret why combining helps more in some settings than others. The market-structure task (Section 3) is open-ended and omission-prone, so heterogeneous personas and checkpoints provide complementary retrieval cues that HuMCal can exploit; the menu-based public-opinion tasks (Section 4) are more variance-sensitive.

C.7. Checkpoint-Upgrade Evaluation (Model-Improvement Compatibility)

Proposition A-7 characterizes a sufficient condition for improvement in the calibration discrepancy. The checkpoint experiment measures realized changes in AR after the candidate set changes. Recalibration without new human data presumes that the original measurement target and human benchmark remain relevant.

We rerun the five-checkpoint (multi-model) market-structure pipeline after upgrading three of the model subgroups from the original batch to newer checkpoints. We replace GPT-4o with GPT-4.1, Grok-3 with

Grok-4, and Mistral-Large-2411 with Mistral-Large-3, while keeping the remaining two checkpoints unchanged. Figure C.5 reports the resulting change (newer checkpoints vs. the original batch) in AR across the same temporal, geographic, and industry evaluations as in Section 3.2.2. AR improves in all transfer evaluations relative to the original checkpoints. Specifically, AR increases by 3.5–9.5% in-sample, 8.0% in temporal transfer, 21.6–39.8% in geographic transfer, and 0.6–9.8% for all three industry targets.

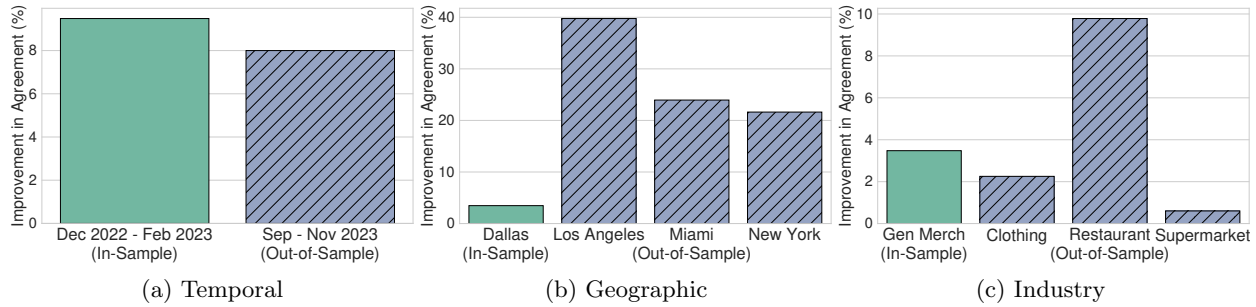


Figure C.5 Checkpoint-upgrade evaluation for the multi-checkpoint market-structure setting. Bars report improvement in agreement relative to the original checkpoint batch, normalized by the original batch's remaining room for improvement. Green bars are source evaluations and blue bars are transfer evaluations. Positive values favor the upgraded batch.

Appendix D: Market-Structure Data and Supplementary Analyses

This section describes how we collected and processed the data and provides supplementary results for the market-structure application.

D.1. LLM Data Collection

This section details the NAICS codes and focal brands used in the prompt (Section 3.1), followed by a list of model versions and checkpoints for the LLMs employed in model-based subgroups.

D.1.1. NAICS Definition

Our study focuses on four industry groups:

- **General Merchandise Retailers (455)**: These stores sell a diverse range of products across multiple categories (food, clothing, household goods, electronics, beauty products, and more) within a single location. They allow customers to find various items in one trip.²⁷

Note:²⁸ Grocery stores, supermarkets, clothing stores, electronics retailers, and home improvement retailers are NOT included in this category.

- **Clothing and Clothing Accessories Retailers (458110)**: These retailers specialize in selling new clothing and fashion accessories, either as general apparel stores or specialty shops.

Note: Stores that primarily sell shoes, precious jewelry and watches, athletic uniforms, or a broad range of products (department stores) across different categories do NOT belong in this category.

- **Limited-Service Restaurants (722513)**: These restaurants require customers to order and pay before receiving their food. Dining options typically include eating on-site, taking food to go, or delivery.

Note: Restaurants that provide full table service (where customers order at the table and pay after eating), cafeterias, grill buffets, buffets, and drinking places do NOT belong in this category.

- **Supermarkets and Other Grocery Retailers (except Convenience Retailers) (445110)**: This category includes supermarkets and grocery stores that primarily sell a full range of food products, including canned and frozen foods, fresh fruits and vegetables, and fresh and prepared meats, fish, and poultry.

Note: Stores that mainly sell fuel, offer a limited selection of products (convenience stores), or carry a wide variety of goods across different categories (department stores and warehouse clubs) do NOT belong in this category.

D.1.2. Focal Brand Selection Within each of the four chosen NAICS categories, we select the top ten brands as focal brands based on their customer count. We fix this brand set using the entire Yodlee dataset (from December 1, 2022, to November 30, 2023, limited to transactions in USD and occurring at physical locations).

D.1.3. LLM Model Versions and Checkpoints Table D.1 lists the specific versions of the LLM families used in our study.

²⁷ The definitions are modified from the official definitions in the 2022 NAICS Manual.

²⁸ To improve the clarity of the prompt for LLMs, we also provide an additional note to differentiate the defined NAICS from similar codes.

Table D.1 Model Version and Update Information on Azure

Model Name	Model Version on Microsoft Azure	Date Updated on Azure
Llama-3.3-70B-Instruct	5	Apr 15, 2025
gpt-4o	2024-11-20	Dec 2, 2024
Mistral-Large-2411	2	Apr 15, 2025
cohere-command-a	1	Apr 15, 2025
grok-3	1	May 19, 2025
Mistral-Large-3	1	Dec 2, 2025
gpt-4.1	2025-04-14	Apr 14, 2025
grok-4	1	Sep 29, 2025

D.2. Processing Human Behavioral Data

We collected bank transactional data spanning from December 1, 2022, to November 30, 2023, from Yodlee²⁹, comprising debit and credit card purchases and payments. This dataset captures instances such as a customer’s purchase at points of interest (POIs), with the transaction details recorded by the financial institution. This section outlines our data processing procedure, beginning with merging Yodlee data with external sources to append desired attributes (e.g., PUMA, NAICS), followed by a detailed description of the data selection criteria.

The licensed Yodlee extract includes a time-varying customer-level field containing information about income class. For each month from December 2022 through November 2023, we use the record with that start date when available and otherwise the record with the nearest start date; we then assign each customer the modal class across the twelve months. Customers without an income-class record are excluded from the income-conditioned benchmarks. We map the seven classes to the income categories listed in Section 3.1. The source metadata do not identify the class as personal or household income or as reported or inferred, so we treat it as a provider-supplied income classification. ACS income is used only to construct population-share weights, not to assign Yodlee customers to income groups.

D.2.1. Data Alignment Our study collects demographic data from the American Community Survey 1-Year Data (ACS-1, 2022) for subgroup split, which uses the Public Use Microdata Area (PUMA) as a basic geographic partition. To align the ZIP codes provided in the Yodlee dataset with the PUMAs used by ACS-1, we used the Geographic Correspondence (Geocorr) 2022 crosswalk file³⁰ from the Missouri Census Data Center. Because the two geographic units are not strictly aligned, wherein a single PUMA contains multiple ZIP codes, while a ZIP code can be split across several PUMAs, co-purchase statistics calculated by ZIP code are scaled by the allocation factor from the crosswalk file.

Geographic allocation can yield equal fractional co-purchase counts, which we break by sorting first on count (descending) and then on brand name (ascending). Neither general-merchandise source cell has a tie among the retained top 10. No tie straddles the tenth-position cutoff, so the rule affects only the order of retained brands, never which brands are retained.

We classify brands based on the North American Industry Classification System (NAICS). At first, each brand’s 2017 NAICS code was obtained from SafeGraph Places³¹ by brand name matching. Then the 2017

²⁹ <https://www.yodlee.com/data-analytics/what-is-transaction-data>

³⁰ <https://mcdc.missouri.edu/applications/geocorr2022.html>

³¹ <https://docs.safegraph.com/docs/places#brand-info>

code is updated to its 2022 version using the “2017 to 2022 change” appendix in the U.S. Census Bureau’s 2022 NAICS Manual³².

Target population shares. For the adjustment in Equation (3), a demographic cell with an unavailable target share must be merged with another cell, newly calibrated, or omitted before renormalizing. We omit such cells in our analysis. In the East Collin County target at the \$60K–\$75K income level, the female and male 18–24 cells have no entries in the 2022 ACS-1 PUMS table used to construct target shares. In the Miami target of the 85-subgroup transfer, the female 18–24 cell at the \$150K+ income level is likewise absent, leaving 83 of the 84 demographic cells.

D.2.2. Data Selection Criteria We construct the analysis sample by restricting the raw Yodlee transaction feed to (i) purchases occurring between December 1, 2022, and November 30, 2023, (ii) transactions denominated in USD and recorded as in-person purchases, and (iii) merchants in the four NAICS categories defined in Appendix D.1.1, which also scope the LLM prompts (Section 3.1).

The analysis sample contains millions of transactions from tens of thousands of customers and more than one thousand merchant brands. To comply with the confidentiality and data-use requirements governing the transaction data, we suppress exact transaction and customer counts by metropolitan area.

Temporal transfer design. The focal-brand universe defined above and the checkpoints in Table D.1 are held fixed across methods and periods. The temporal analysis therefore tests whether weights estimated against the earlier human benchmark remain useful against the later benchmark under this fixed brand universe and checkpoint set.

D.3. Objective Function for Regional Market Structure (Ordinal Co-Purchase Data)

This appendix describes Steps 4 and 5 of HuMCal’s roadmap from Figure 1 for the regional market structure task introduced in Section 3. Steps 1–3 defined the question, chose subgroups, and collected data; we now show how those data are combined.

D.3.1. Data recap (output of Step 3). For every focal brand $q \in \{1, \dots, Q\}$, let $B_q^H = (b_{q,1}^H, \dots, b_{q,M_q}^H)$ denote the human co-purchase benchmark in decreasing order, where $M_q := |B_q^H| \leq M$. For each subgroup g and each pair $1 \leq j < k \leq M_q$, define the concordance sign $s_{q,g,jk} \in \{+1, -1, 0\}$: +1 if subgroup g ’s retained LLM list of length K_{qg} contains both $b_{q,j}^H$ and $b_{q,k}^H$ and ranks $b_{q,j}^H$ higher, −1 if it contains both and ranks $b_{q,k}^H$ higher, and 0 if either brand is absent from that list. Here K_{qg} is the retained list length, with $17 \leq K_{qg} \leq 20$ (Section 3.1). The weighted concordance margin

$$d_{q,jk}(\mathbf{w}) = \sum_{g=1}^G w_g s_{q,g,jk}, \quad 1 \leq j < k \leq M_q,$$

is affine in $\mathbf{w}$; positive values indicate weighted agreement with the human ordering.

Weighted pairwise majorities can cycle. Calibration and AR use the margins $d_{q,jk}(\mathbf{w})$ pair by pair; neither requires a total ranking. The hierarchical extension converts margins to a ranking using the rule in Appendix D.6.

³²https://www.census.gov/naics/reference_files_tools/2022_NAICS_Manual.pdf

D.3.2. Discrepancy function (Step 4). Because both outputs are ordinal lists, we adopt the pairwise logistic loss from Section 2.2.1:

$$\ell_{q,jk}(\mathbf{w}) = \log(1 + \exp[-d_{q,jk}(\mathbf{w})]), \quad 1 \leq j < k \leq M_q,$$

where the human ordering is encoded in the sign convention of $s_{q,g,jk}$, so positive margins indicate weighted agreement with the human order. Pairs covered by no subgroup contribute the constant $\log 2$. Omitting those terms and averaging the retained losses preserves the exact minimizer of the unregularized full-pair objective below.

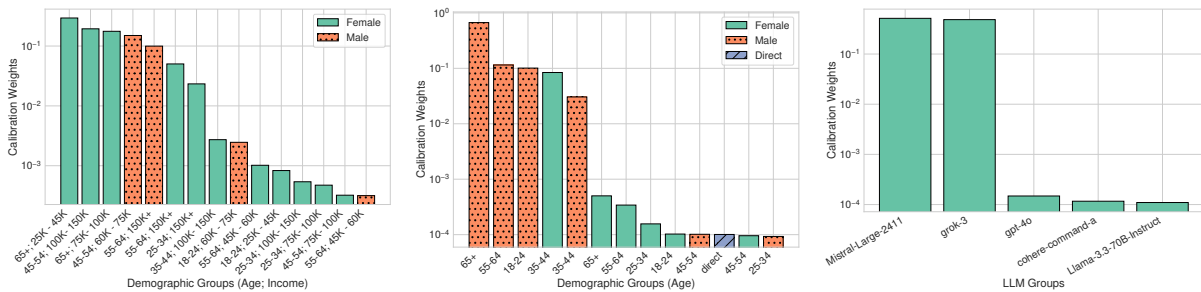
D.3.3. HuMCal optimization (Step 5). Plugging $\ell_{q,jk}$ into the generic program (HuMCal-OPT) yields

$$\mathbf{w}^* \in \arg \min_{\mathbf{w} \in \Delta_G} f_{\text{MS}}(\mathbf{w}), \quad f_{\text{MS}}(\mathbf{w}) = \sum_{q=1}^Q \sum_{1 \leq j < k \leq M_q} \ell_{q,jk}(\mathbf{w}). \quad (\text{HuMCal-OPT}^{\text{MS}})$$

Each $d_{q,jk}(\mathbf{w})$ is affine in $\mathbf{w}$, and each $\ell_{q,jk}$ is convex in $d_{q,jk}$, so f_{MS} is convex on the simplex $\Delta_G = \{\mathbf{w} \geq 0, \sum_g w_g = 1\}$. Hence any local minimizer is global by Theorem A-1. As with the general results, strict convexity and uniqueness are not automatic with these surrogates and require additional injectivity or regularization conditions. Theorem A-1 and Corollary A-3 apply to this exact closed-simplex program. The market-structure transfer panels evaluate weighted margins only on target human-listed pairs and their reference ordering under AR. They therefore test reuse of source weights, not benchmark-free recovery of a target top- M list or total ranking.

D.3.4. Interpretation. The calibrated weights w_g^* measure how much each subgroup contributes to the mixture: subgroups whose LLM rankings agree closely with observed co-purchase behavior receive larger weights, while subgroups that disagree are down-weighted toward zero.

D.4. Calibration Weights for Market Structure Learning



(a) Dallas metro, top 15 of 85. (b) Park Cities (Dallas), \$60K–\$75K. (c) Dallas metro, five checkpoints.
Figure D.1 Calibrated LLM ensemble weights across different aggregate granularities and subgroup definitions.

In the Dallas-metro-level calibration with 85 subgroups (Figure D.1a), two female subgroups (i.e., women aged 65+ with incomes between \$25K and \$45K, and aged 45–54 with incomes between \$100K and \$150K) receive the two highest weights, approximately 29.4% and 19.4%, respectively. In contrast, the bottom eight of the top-15 subgroups contribute less than 1% in total. For Dallas as a whole, the calibrated mixture places the most weight on prompts representing these two female demographic groups.

We then shift to the micro-market calibration with 13 subgroups (Figure D.1b). Within the Park Cities (Dallas)×\$60K–\$75K cell, the highest weights shift to *male* segments, particularly men aged 65+; the top four cells absorb close to 96.8% of the total mass.

When subgroups are defined by model checkpoints rather than demographics (Figure D.1c), the mass is again concentrated: the two leading checkpoints account for 99.96% of the ensemble weight. The final three checkpoints receive only trace weights. Under the softmax fit a checkpoint with negligible weight still contributes the pairs that only it covers to the weighted vote, which is why randomizing the third checkpoint’s ranking still lowers agreement in Appendix C.5.2 (see Appendix A.6).

Across all three experiments, HuMCal allocates much of the weight to a small subset of demographic or model segments, with near-zero mass assigned to many weak or redundant subgroups. This concentration is consistent with uneven subgroup quality and redundancy in these datasets. This is an empirical pattern. Simplex-constrained convex objectives can have either dense interior optima or boundary solutions.

A practical implication is that if certain subgroups receive near-zero weight across related calibration tasks, they are natural candidates for pruning in later data collection, after checking that the omission does not lower agreement.

D.5. Agreement Rate (AR) for Market Structure Learning

Figures D.2–D.5 report the agreement rates for market structure learning, complementing the performance improvements discussed in Section 3.2.³³

Pair coverage and ordering. AR reflects both pair coverage and ordering; ties are nonconcordant. In the in-sample checkpoint comparison (Figure D.2(c)), uniform averaging and the implemented HuMCal fit give positive weight to the same five checkpoint outputs, so both cover 95.6% of human-listed pairs. Their raw 6.9-percentage-point AR difference therefore comes from the weighted votes on that common covered-pair set. Relative to a single checkpoint, retaining multiple outputs also increases coverage because different checkpoints omit different pairs.

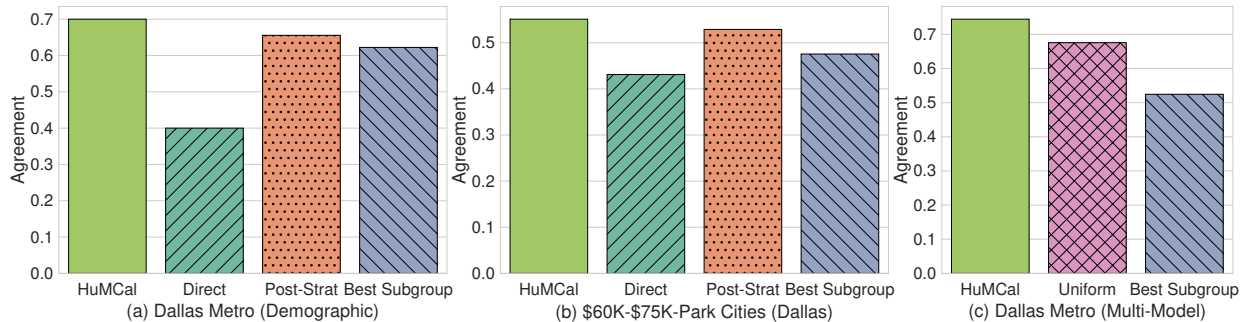


Figure D.2 In-Sample Performance Evaluation Measured by AR (December 2022–November 2023).

³³ For a single list, $\mathbf{w} = \mathbf{e}_g$ assigns all weight to subgroup g . Metro benchmarks have $M_q = 10$; sparse-cell benchmarks can have $M_q < M$, so Equation (4) uses each benchmark’s actual number of pairs.

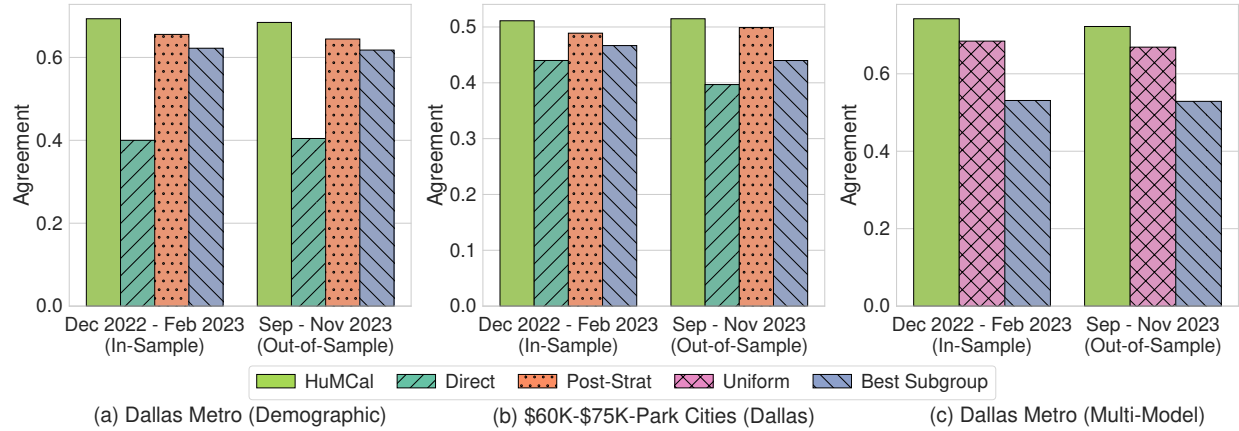


Figure D.3 Temporal Transferability (In-Sample and Out-of-Sample) Evaluated by AR.

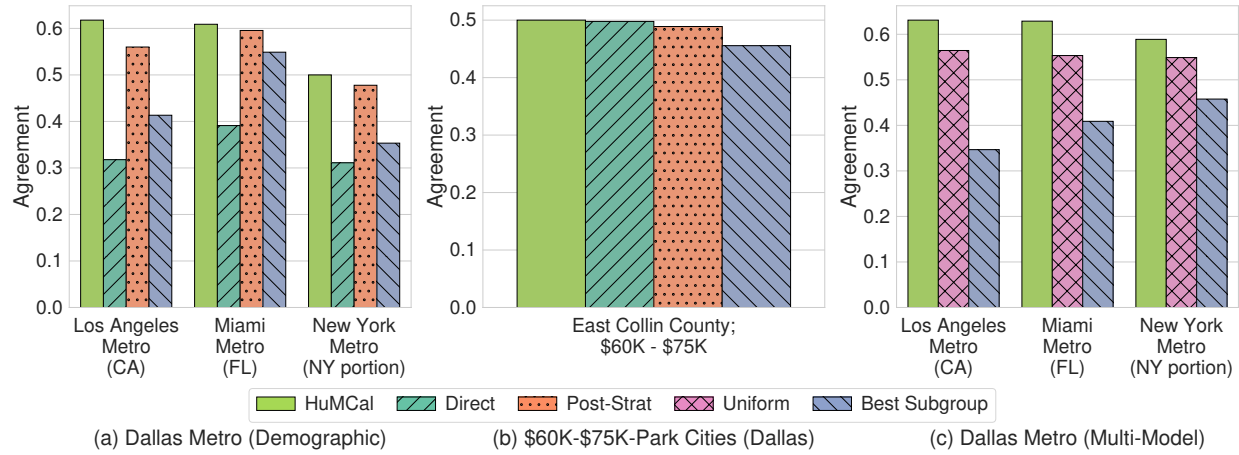


Figure D.4 Geographic Transferability (Out-of-Sample) Evaluated by AR. In panel (b), post-stratification uses the target region's demographic shares.

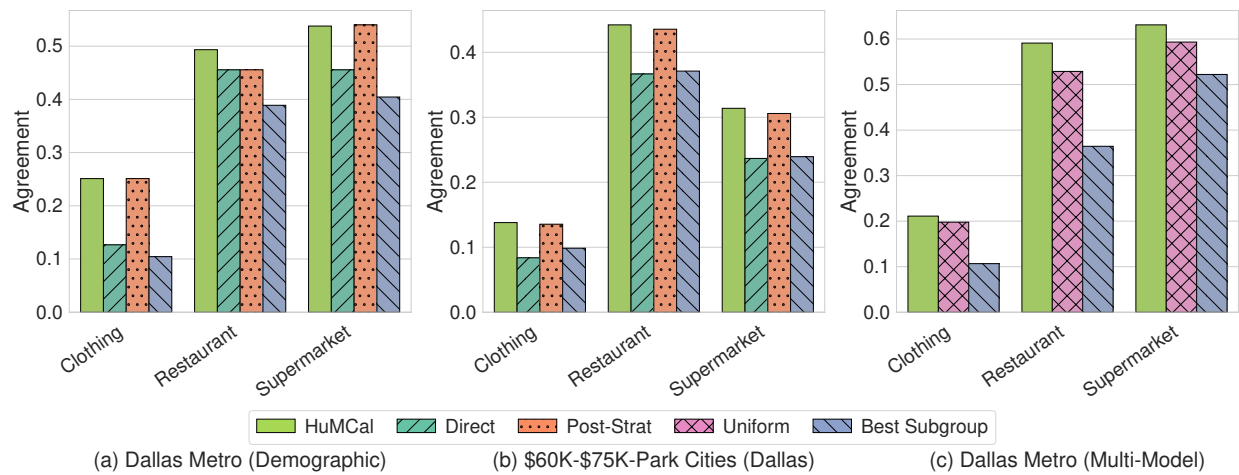


Figure D.5 Industry Transferability (Out-of-Sample) Evaluated by AR.

D.6. Hierarchical Calibration with Subgroup-Level Benchmarks

When subgroup-level human benchmarks are available, we can adopt the two-stage hierarchical calibration framework introduced in Section 2.2.2.

Stage 1: Local subgroup calibration. For group g , suppose K_g “sub-personas” of the LLM yield outputs $\mathbf{z}_{q,g,i}^L$ for $i = 1, \dots, K_g$. One solves:

$$\begin{aligned} \min_{\mathbf{w}_g} \quad & \sum_{q=1}^{K_g} \mathcal{D}(\hat{\mathbf{z}}_{q,g}^L(\mathbf{w}_g), \mathbf{z}_{q,g}^H), \\ \text{s.t.} \quad & w_{g,i} \geq 0, \quad \sum_{i=1}^{K_g} w_{g,i} = 1, \end{aligned} \tag{D.1}$$

where $\hat{\mathbf{z}}_{q,g}^L(\mathbf{w}_g) = \sum_{i=1}^{K_g} w_{g,i} \mathbf{z}_{q,g,i}^L$. Solving (D.1) yields an optimal local weight vector $\mathbf{w}_g^*$, bringing group g ’s LLM outputs into agreement with its own human data.

Before Stage 2, we convert the Stage 1 pairwise margins into a total ranking for each focal-brand question and primary subgroup.³⁴

1. Each pair with a nonzero margin contributes one directed edge, oriented by the sign of the margin and weighted by its magnitude. A pair whose margin is zero is a tie and contributes no edge.
2. Apply a topological sort to the resulting directed graph, emitting alphabetically among the brands currently subject to no unbroken incoming edge.
3. If the topological sort stalls before all brands have been ranked, remove a minimum-weight edge between two unranked brands and resume the sort. The selected edge need not lie on the cycle that stalled the sort. Each deletion removes one of finitely many edges, so the procedure terminates.³⁵

Stage 2: Global recalibration. The resulting rankings are treated as fixed subgroup outputs in Stage 2. We apply the original HuMCal program in Equation (HuMCal-OPT) to estimate global calibration weights across those rankings and form the aggregate estimate. Stage 2’s convexity and fixed-weight guarantees apply conditional on the nonlinearly projected Stage 1 rankings; the two weight vectors are not jointly optimized. In our empirical study within the Dallas metro area, we define primary subgroups by income brackets, further segmented into twelve detailed age×gender combinations. Each Stage 1 fit therefore has 13 candidates: the twelve age×gender sub-personas plus that income bracket’s own aggregate prompt.

At the first stage, in-sample agreement rate increases compared to direct elicitation (from 31.4% to 46.1%, mean 38.0%) and post-stratification (from 0.0% to 11.7%, mean 5.9%). Out-of-period (transferring from December 2022–February 2023 to September–November 2023), improvements similarly range from 30.0% to 43.6% (mean 36.4%) relative to direct elicitation, and from 2.2% to 10.3% (mean 5.9%) relative to post-stratification (Figure D.6).

³⁴ For the empirical projection, the Stage 1 weights are estimated from the human-benchmark pairs and then applied to all pairs appearing in the candidate rankings for the same focal-brand question, including pairs absent from the human benchmark.

³⁵ If a cycle admits no further deletable edge, the brands still unranked are appended alphabetically, as are any brands absent from the final order.

A second-stage calibration further enhances predictive agreement rate. In-sample improvements reach 53.0% over direct elicitation and 18.1% relative to post-stratification, while out-of-sample gains are 50.4% and 16.9%, respectively (Figures D.6b and D.7b). Panel (b) of both figures compares the complete hierarchy with “One-Stage Global HuMCal,” fitted over 84 retained adult age $\times$ gender $\times$ income personas plus the aggregate prompt, and “HuMCal + Post-Strat (Ablation),” which weights the seven fixed Stage 1 income-level rankings by population shares. The former changes both candidate outputs and benchmark information; the latter isolates Stage 2. On that comparison the second stage raises metro-level AR by 14.8% in sample and 13.6% out of sample.

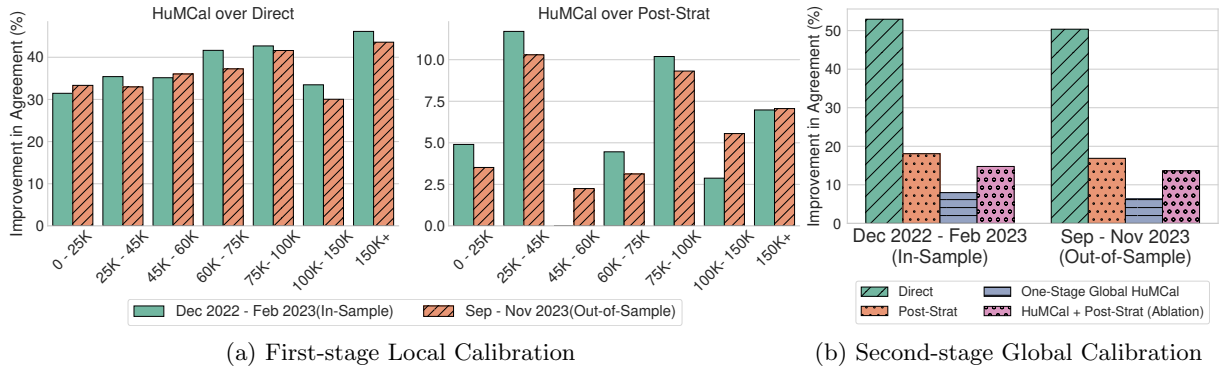


Figure D.6 Hierarchical Calibration: Improvements in AR (In-Sample and Out-of-Sample).

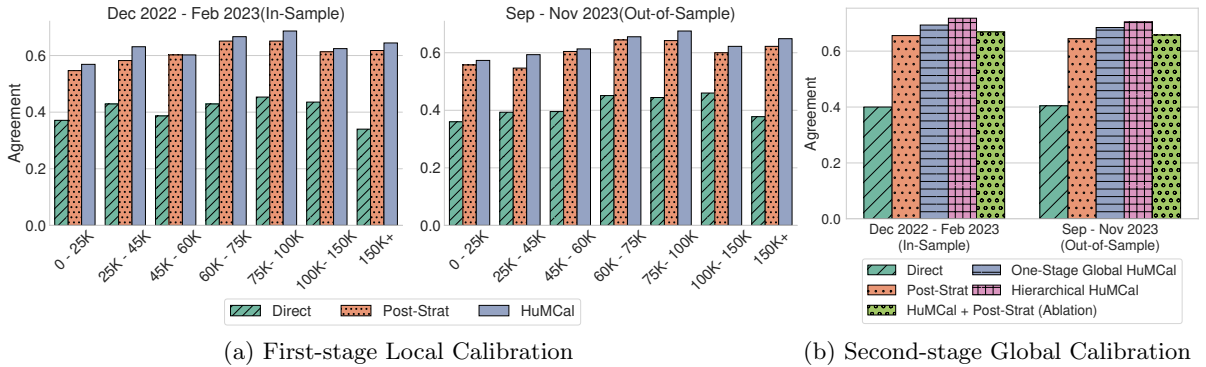


Figure D.7 Hierarchical Calibration: AR Levels (In-Sample and Out-of-Sample).

D.7. Does HuMCal Still Help When No Persona Beats the Aggregate Prompt?

Stromberg et al. (2025) document settings—specifically, survey tasks with discrete response formats such as multiple-choice and Likert scales—where a single aggregate prompt (“direct elicitation”) outperforms post-stratifying subgroups. One might therefore question whether HuMCal adds value when no retained persona outperforms the aggregate prompt. Although Corollary A-3(b) guarantees weak dominance over direct elicitation in the calibration loss when the aggregate prompt is included as a subgroup, it does not

by itself imply weak dominance in AR. We therefore evaluate AR empirically in this deliberately challenging setting.

We work in the \$60K–\$75K income tier of Park Cities (Dallas) using 12 gender-by-age personas plus one direct-elicitation subgroup. To create a deliberately challenging pool in AR terms, we discard every persona whose raw AR with human aggregate data in the general-merchandise source setting exceeds direct elicitation. This leaves three personas plus the aggregate prompt for calibration in the industry and spatial analyses (Figure D.8a–D.8b). One of the three ties direct elicitation. Even with this restricted candidate pool, HuMCal lifts in-sample AR by 9.4% and, when transferred to clothing, restaurant, and supermarket tasks, adds 4.2–8.4% over direct elicitation; the gain is 4.4% in a neighboring county.

Further, a temporal test applying the same screen in the earlier three-month source window (eight weak personas together with direct elicitation) improves AR by 9.1% in that window and by 18.0% when the calibrated weights are carried forward to September–November 2023 (Figure D.8c), in both cases over direct elicitation. Thus HuMCal surpasses direct elicitation even when no retained persona outperforms it individually in the source setting.

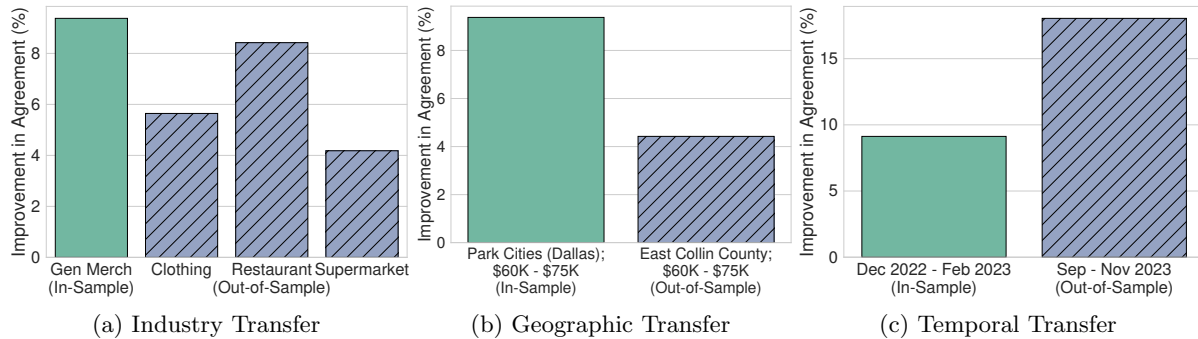


Figure D.8 HuMCal trained on personas that do not outperform direct elicitation in the source setting: in-sample gains and transfers across industry, county, and time. The y-axis reports the percent of direct elicitation’s remaining disagreement that HuMCal closes.

Appendix E: Public Opinion Data and Supplementary Analyses

This appendix describes the data and provides supplementary analyses for Section 4.

E.1. Data Processing on Public Opinion Surveys

Human responses come from the Pew Research Center’s American Trends Panel. LLM responses come from Santurkar et al. (2023), who collected responses from nine checkpoints developed by AI21 Labs and OpenAI.³⁶

The human dataset included responses from multiple Pew surveys covering a variety of topics. Each survey contained a set of multiple-choice questions and responses from thousands of respondents across the United States, along with their demographic information. This panel consists of approximately 10,000 participants recruited over multiple years, with many individuals taking the survey repeatedly. Each year, a subset of panelists is invited to participate in the survey.

We use the category order in the prepared OpinionsQA inputs as a one-dimensional support and code the categories as consecutive integers $1, \dots, N_q$; for bipolar items with an explicit midpoint, categories are ordered from one directional endpoint, through the midpoint, to the opposite endpoint. The prepared distributions omit the “Refused” category and renormalize the retained human and model probabilities. For 11 questions, they also omit a “Not sure” category and renormalize over the remaining responses. We retain 372 preference questions in clustering and evaluate 370 of them, whose answer choices form an ordinary ordinal or bipolar scale. We exclude two questions from evaluation, because one has unordered answer choices, and the other asks the model about being white but human respondents about their own racial background. In Appendix E.1.1, we identify different question clusters associated with these questions.

E.1.1. Clustering Public Opinion Survey Questions We encode each pair comprising a question identifier and its retained question record (including the question text and answer options) with the pre-trained sentence-transformer `all-mpnet-base-v2`. We apply Ward hierarchical clustering to the resulting embeddings and cut the dendrogram at Ward linkage height 3, yielding five clusters. Because each question identifier includes a survey-wave suffix, the clustering input contains both question content and wave information. The descriptively labeled gender, race, and gun clusters are concentrated in waves 36 (65 of 79 questions), 43 (38 of 45), and 26 (21 of 22), respectively. The analysis examines transfer across question clusters that can differ in both topic and survey wave.³⁷ Each source cluster supplies one fitted weight vector, which is reused for all of its off-diagonal target-cluster comparisons.

The visualizations in Figure E.1 include a hierarchical clustering dendrogram (Figure E.1a), the number of questions per cluster (Figure E.1b), a two-dimensional t-distributed stochastic neighbor embedding (t-SNE) projection of the question embeddings (Figure E.1c), and distances between cluster centroids (Figure E.1d). We assign each cluster a label based on word frequency and present ten example questions for each cluster in Appendix E.1.2. As can be seen from Figures E.1c and E.1d, clusters 4 (**economy**) and 5 (**research**) are closer to other clusters, while cluster 3 (**gun**) is farther from the others, especially from clusters 1 and 2.

³⁶The LLMs they investigated include `j1-grande`, `j1-grande-v2-beta`, `davinci`, `text-davinci-001`, `text-davinci-002`, `ada`, `j1-jumbo`, `text-ada-001`, `text-davinci-003`.

³⁷Economy and research separate least cleanly of the five clusters. Their silhouette widths are 0.07 and 0.03, against 0.14 for gender, 0.20 for race, and 0.39 for gun. We read that one split as approximate.

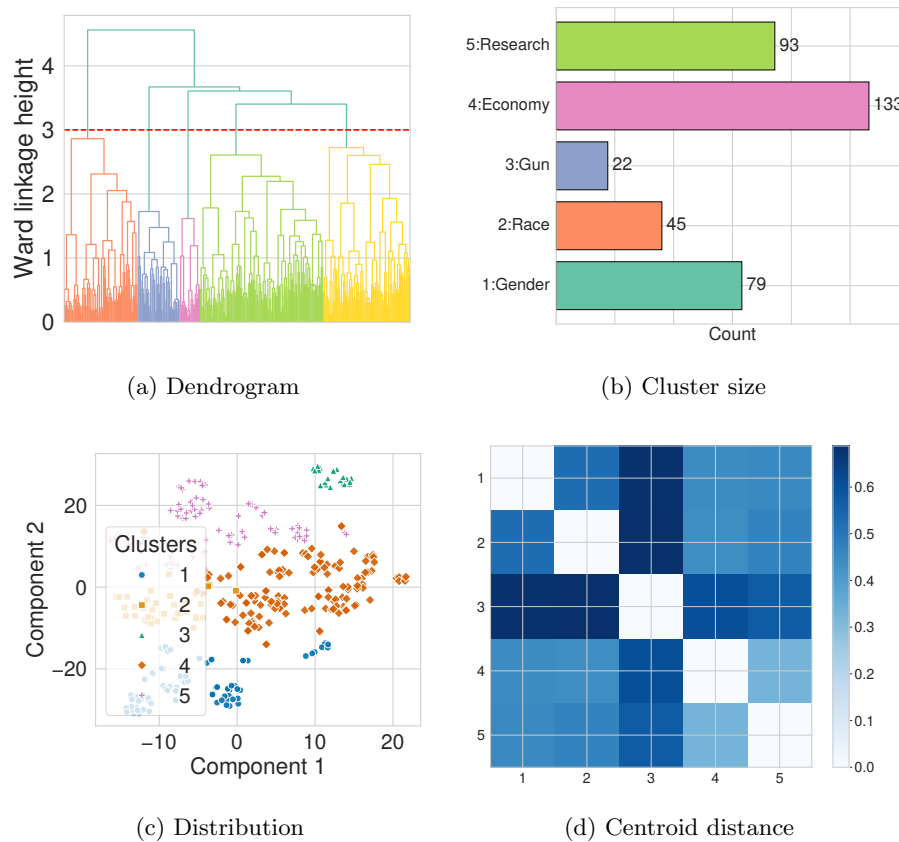


Figure E.1 Visualization of public opinion question clusters: dendrogram (a), cluster sizes (b), t-SNE distribution (c), and centroid distances (d).

E.1.2. Example Questions in Each Question Cluster In this appendix, we present ten example questions for each cluster. These examples are included to illustrate topical coverage; for the WD-based analysis, each retained item is coded only after verifying that its response categories admit a defensible total order, with bipolar items coded from one endpoint, through the midpoint, to the opposite endpoint.

Cluster 1: Gender

- In general, do you think men or women in top executive business positions are better at working well under pressure?
- In general, do you think men or women in top executive business positions are better at being willing to take risks?
- In general, do you think men or women in top executive business positions are better at negotiating profitable deals?
- In general, do you think men or women would do a better job running a large bank or financial institution types of companies?
- In general, do you think men or women would do a better job running a professional sports team types of companies?

- In general, do you think men or women in high political offices are better at being compassionate and empathetic?
- In general, do you think men or women in high political offices are better at being willing to take risks?
- In general, do you think men or women in high political offices are better at serving as a role model for children?
- And thinking about some specific policy areas, in general, do you think men or women in high political offices are better at handling economic conditions?
- And thinking about some specific policy areas, in general, do you think men or women in high political offices are better at dealing with immigration policy?

Cluster 2: Race

- How likely, if at all, is it that black people in our country will eventually have equal rights with whites?
- Since President Trump was elected, do you think it has become more common or less common for people to express racist or racially insensitive views, or is it about as common as it was before?
- Since President Trump was elected, do you think it has become more acceptable or less acceptable for people to express racist or racially insensitive views, or is it about as acceptable as it was before?
- How much, if at all, do you think the legacy of slavery affects the position of black people in American society today?
- Thinking again about race and race relations in the U.S. in general, how well, if at all, do you think each of these groups get along with each other in our society these days? Blacks and Hispanics
- Thinking again about race and race relations in the U.S. in general, how well, if at all, do you think each of these groups get along with each other in our society these days? Hispanics and Asians
- Do you think each is a major reason, minor reason, or not a reason why black people in our country may have a harder time getting ahead than white people? Racial discrimination
- Do you think each is a major reason, minor reason, or not a reason why black people in our country may have a harder time getting ahead than white people? Less access to good quality schools
- Do you think each is a major reason, minor reason, or not a reason why black people in our country may have a harder time getting ahead than white people? Less access to high-paying jobs
- Would you say that black people are treated less fairly than white people, white people are treated less fairly than black people, or both are treated about equally in hiring, pay and promotions situations?

Cluster 3: Gun

- Would you say the following was a reason or was not a reason why there were guns in your household when you were growing up? For protection?
- Would you say the following was a reason or was not a reason why there were guns in your household when you were growing up? For hunting?

- Would you say the following was a reason or was not a reason why there were guns in your household when you were growing up? For sport shooting, including target shooting and trap and skeet?
- How much, if at all, do you think the ease with which people can legally obtain guns contributes to gun violence in the country today?
- How much, if at all, do you think the amount of gun violence in movies and television contributes to gun violence in the country today?
- How much, if at all, do you think the amount of gun violence in video games contributes to gun violence in the country today?
- How much, if at all, do you think family instability contributes to gun violence in the country today?
- Thinking about gun owners who have children in their home, how important do you think it is for them to: Advise visitors that there are guns in the house?
- Thinking about gun owners who have children in their home, how important do you think it is for them to: Advise visitors with children that there are guns in the house?
- Thinking about gun owners who do not have children in their home how important do you think it is for them to: Keep all of their guns in a locked place?

Cluster 4: Economy

- How much power and influence do you think people who are wealthy have on today's economy?
- Thinking about the nation's economy, how would you rate economic conditions in this country today?
- How much, if at all, do you think the cost of health care is contributing to your opinion about how the economy is doing?
- Do you think the country's current economic conditions are helping or hurting people who are wealthy?
- Do you think the country's current economic conditions are helping or hurting the middle class?
- Do you think the country's current economic conditions are helping or hurting people who are poor?
- Do you think the country's current economic conditions are helping or hurting older adults?
- Do you think the country's current economic conditions are helping or hurting young adults?
- Do you think the country's current economic conditions are helping or hurting people who are white?
- Do you think the country's current economic conditions are helping or hurting people who are black?

Cluster 5: Research

- Thinking about medical research scientists, how often would you say they care about the best interests of the public?
- Thinking about environmental research scientists, how often would you say they do a good job conducting research?
- Thinking about environmental research scientists, how often would you say they admit mistakes and take responsibility for them?

- Thinking about environmental research scientists, how often would you say they are transparent about potential conflicts of interest with industry groups in their research?
- Thinking about environmental research scientists, how often would you say they care about the best interests of the public?
- Thinking about nutrition research scientists, how often would you say they do a good job conducting research?
- Thinking about nutrition research scientists, how often would you say they provide fair and accurate information when making statements about their research?
- Thinking about nutrition research scientists, how often would you say they are transparent about potential conflicts of interest with industry groups in their research?
- Overall, do you think research misconduct by environmental research scientists is?
- Overall, do you think research misconduct by nutrition research scientists is?

E.2. Objective Function for Public-Opinion Learning (Categorical Data)

This appendix spells out Steps 4 and 5 of HuMCal’s roadmap from Figure 1 for the public-opinion calibration task introduced in Section 4. Steps 1–3 defined the question, chose subgroups, and collected data; we now show how those data are combined.

E.2.1. Discrepancy function (Step 4). Because all retained answers are discrete over response categories with a defensible total order, we measure discrepancy with the one-dimensional Wasserstein distance:

$$\mathcal{D}_{\text{WD}}(\mathcal{Z}_q^L(\mathbf{w}), \mathcal{Z}_q^H) = \sum_{i=1}^{N_q-1} \left| F_L^{(q)}(i; \mathbf{w}) - F_H^{(q)}(i) \right|, \quad F_L^{(q)}(i; \mathbf{w}) = \sum_{g=1}^G w_g F_{L,g}^{(q)}(i),$$

where $F_{L,g}^{(q)}(i)$ is subgroup g ’s LLM cumulative distribution function (CDF), and $F_H^{(q)}(i)$ is the aggregate human-benchmark CDF, each evaluated at category i .

E.2.2. HuMCal optimization (Step 5). Plugging the Wasserstein loss into the generic HuMCal program (Equation (HuMCal-OPT)) yields the convex objective

$$\mathbf{w}^* \in \arg \min_{\mathbf{w} \in \Delta_G} f_{\text{OP}}(\mathbf{w}), \quad f_{\text{OP}}(\mathbf{w}) = \sum_{q=1}^Q \sum_{i=1}^{N_q-1} \left| F_L^{(q)}(i; \mathbf{w}) - F_H^{(q)}(i) \right|. \quad (\text{HuMCal-OPT}^{\text{OP}})$$

Here $\Delta_G = \{\mathbf{w} \geq 0 \mid \sum_g w_g = 1\}$ is the probability simplex. Each $F_L^{(q)}(i; \mathbf{w})$ is affine in $\mathbf{w}$ and absolute value is convex, so f_{OP} satisfies the conditions of Theorem A-1: the program is convex over Δ_G and at least one global minimizer exists. Uniqueness is not guaranteed (absolute value is not strictly convex in general), but is not required for baseline-safety.

In the main text, we report the normalized agreement score in Equation (5) for interpretability; calibration minimizes the unnormalized WD objective. Accordingly, Proposition A-2 and the associated convexity statements apply directly to this WD objective, not to the normalized agreement score. Because the objective sums absolute CDF gaps over $i = 1, \dots, N_q - 1$, questions with more ordered categories contribute more threshold terms to the calibration loss. Because the implementation codes ordered response categories as consecutive integers $1, \dots, N_q$, the normalization in Equation (5) is $N_q - 1$.

Baseline weighting and selection rules. Let π_g denote persona g 's share in the OpinionsQA data, renormalized over the modeled personas. Population-share weighting divides this share equally among the nine checkpoints, assigning weight $\pi_g/9$ to cell (m, g) (checkpoint m , persona g). The shares are 0.499 and 0.501 for Democrat and Republican, and 0.367, 0.255, and 0.378 for conservative, liberal, and moderate; the same shares apply to every question. We use these proportions as supplied rather than re-estimating them from respondent-level survey weights. Uniform averaging assigns $1/G$ to every cell. The agreement-selected best single subgroup is the cell whose own output attains the highest normalized agreement score (Equation (5)) at unit weight. Under the within-cluster design of Section 4.2.2, it is reselected on the training questions of each split and then scored on that split's held-out questions. For cross-cluster transfer, it is selected on the source cluster alone and scored on the disjoint target cluster.

E.3. Calibrated Weights over Checkpoint–Persona Subgroups

Figure E.2 shows the calibrated weights over checkpoint–persona subgroups, by question cluster.

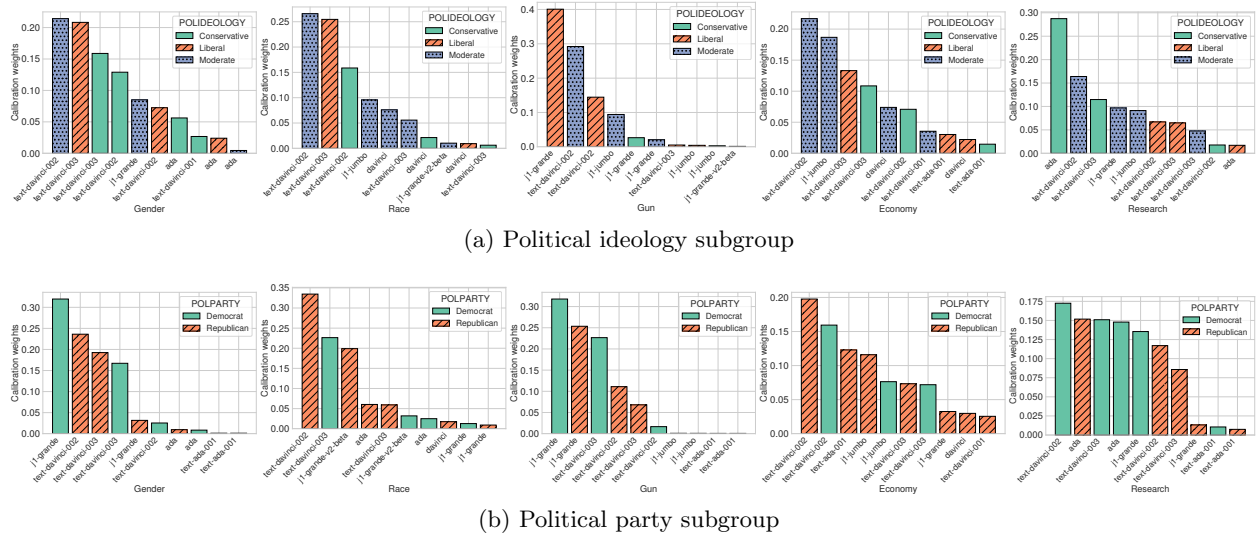


Figure E.2 LLM sample weights calibrated from the public opinion survey: the ten highest-weight checkpoint–persona cells per question cluster. Bars labeled by checkpoint and colored by persona. Panels are labeled by descriptive content label.

E.4. Supplementary Analyses Using Wasserstein Distance

This appendix complements the main results in Section 4.2. We evaluate the same calibrated weights under WD without recalibration.

E.4.1. In-Sample Performance Evaluation (Step 6). Figure E.3 reports the reduction in WD obtained by HuMCaI on the calibration sample itself. Across all five question clusters and both political-persona panels shown, HuMCaI surpasses all three baselines.

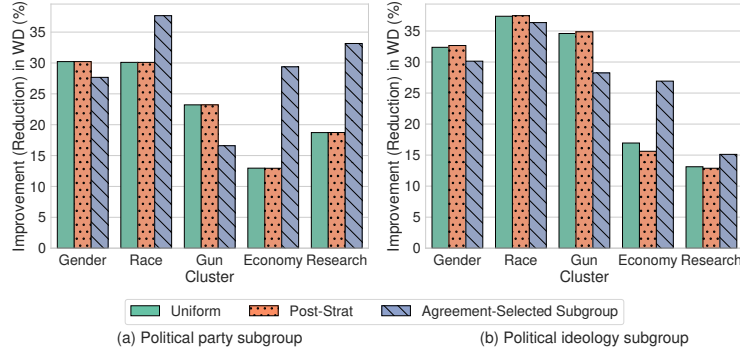


Figure E.3 Relative reduction in Wasserstein distance achieved by HuMCal versus the uniform, post-stratification, and agreement-selected subgroup baselines on the calibration (in-sample) data.

E.4.2. Out-of-Sample Generalization Within and Across Question Clusters (Step 7). To assess generalization, we consider two evaluation settings (Figure E.4). For *within-cluster* generalization (diagonal cells), each question cluster is split 75%/25%; the split is repeated 500 times, and performance is averaged on the held-out subset. For *cross-cluster* transfer (off-diagonals), the model is trained on the entire source cluster and tested on the entire, disjoint target cluster.

Within-cluster. HuMCal reduces WD on held-out questions in 29 of the 30 within-cluster baseline comparisons. The only deficit is the political-party **gun** cluster relative to the agreement-selected best-single-subgroup benchmark (-2.22%); reductions in all other comparisons range from 4.39% to 35.9% .

Cross-cluster. Across 120 source–target–baseline comparisons (20 ordered cluster pairs $\times 3$ baselines $\times 2$ persona settings), HuMCal reduces WD in 113. The 7 increases are modest (all smaller than 4.5%) and concentrate in the political-ideology panel, most often when **research** is the target cluster.

These supplementary analyses reinforce the main findings: HuMCal reduces WD in-sample and in most out-of-sample comparisons. The exceptions suggest that practitioners should test transfer empirically when applying HuMCal to new question clusters or persona settings.

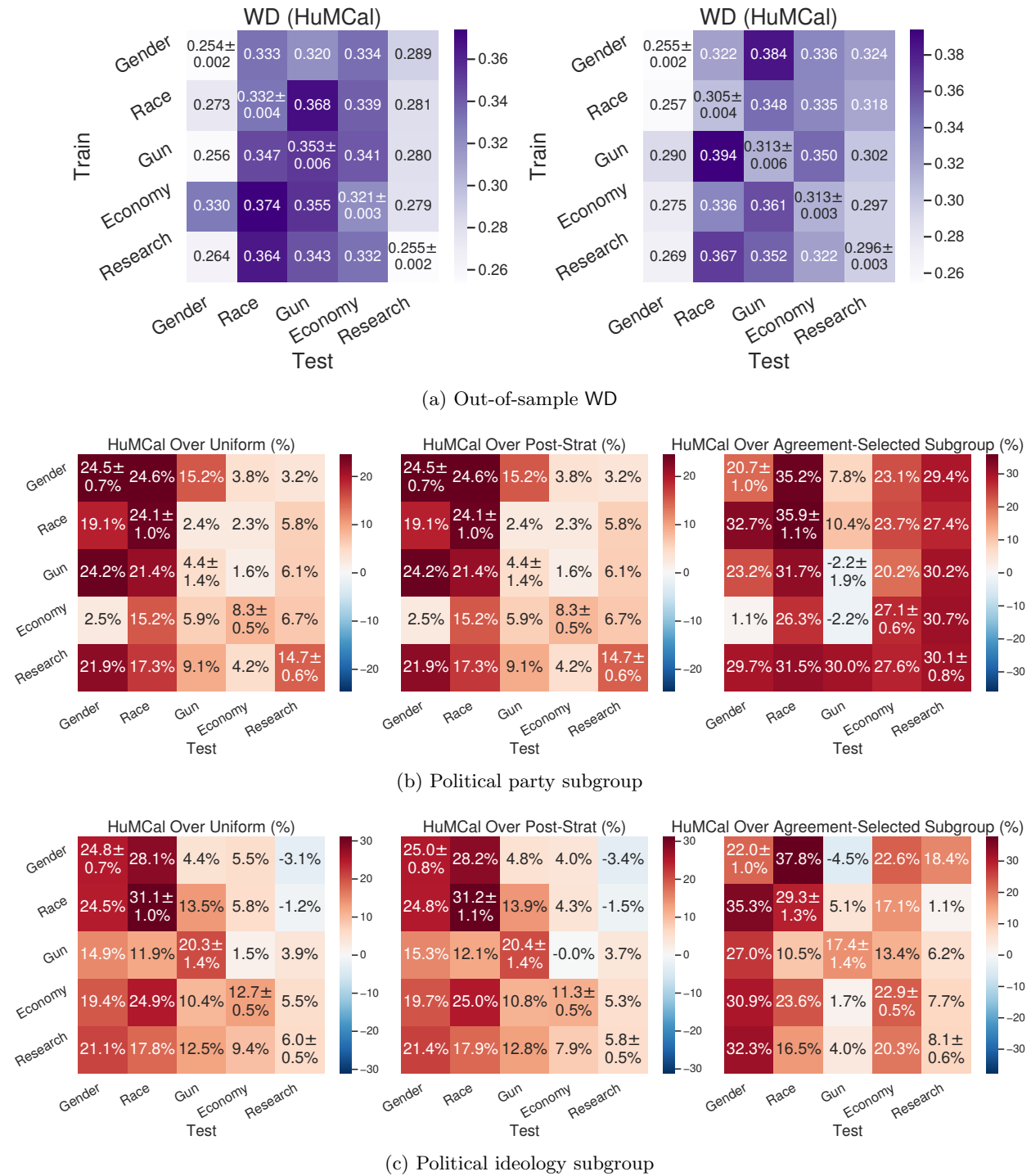


Figure E.4 Transferability of calibration weights across empirically constructed question clusters, which can differ in content and survey wave. Panel (a): out-of-sample WD from HuMCal. In each heatmap, the y-axis is the source cluster and the x-axis is the target cluster. Panels (b) and (c): proportional reduction in WD relative to baselines; red indicates a reduction (better) and blue an increase (worse).

References

- Santurkar S, Durmus E, Ladhak F, Lee C, Liang P, Hashimoto T (2023) Whose opinions do language models reflect? *International Conference on Machine Learning*, 29971–30004 (PMLR).
- Schoenegger P, Tuminauskaite I, Park PS, Bastos RVS, Tetlock PE (2024) Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances* 10(45):eadp1528.
- Stromberg M, Moe WW, Reutterer T, Schweidel DA (2025) Blind spots in broad strokes: Caveats for the use of LLMs in marketing research. Available at SSRN 5145114.